Science AAAS

Supplementary Materials for

Global analysis of protein folding using massively parallel design, synthesis, and testing

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford,
Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran,
Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, David Baker*

*correspondence to: dabaker@u.washington.edu

This PDF file includes:

Methods
Figures S1 to S12
Tables S1 to S3
Definition of scoring metrics
Explanation of supplementary datasets

Other Supplementary Materials for this manuscript includes the following:

Supplementary datasets as archives:
design_pdbs.tar.gz
counts_and_ec50s.tar.gz
unfolded_state_model_params
fig1_thermodynamic_data.csv
design_structural_metrics.tar.gz
design_scripts.tar.gz
stability_scores.tar.gz
protein_and_dna_sequences.tar.gz

Methods

Protein design

All protein design in this work was performed in three stages: (1) backbone construction, (2) sequence design, and (3) selection of designs for testing. Backbone construction (the *de novo* creation of a compact, three-dimensional backbone with a pre-specified secondary structure) was performed using a blueprint-based approach described previously (*34*, *54*). Briefly, blueprint files were built by hand for each topology in order to define (a) the secondary structure at each residue position for that topology, and (b) the strand pairing and register of any β-sheets. All blueprint files are provided along with all design scripts in the data archive design_scripts.tar.gz. These blueprint files were then used to select short three-dimensional fragments from protein crystal structures matching the proposed secondary structure in the blueprint (200 fragments for every 3- and 9-residues-length stretch of the blueprint). Finally, these fragments were assembled into a full protein backbone using Monte Carlo sampling with a coarse-grained energy function (*17*) (including constraints on the hydrogen-bonding pairs of residues as specified for the β-sheets in the blueprint) until the overall backbone matched the specified secondary structure and topology, satisfied compactness criteria, and avoided steric clashes. The same four HHH blueprints (each 43 residues in length) were used for all four design rounds. One 40-residue-length EHEE blueprint was used for all four design rounds. A total of 42, 12, 27, and 7 HEEH blueprints (each 43 residues in length) were used in design rounds 1-4 respectively. Blueprints for each round were selected based on the stabilities of designs from the prior round; new blueprints were also introduced in design rounds 2 and 3. A total of 2, 1, 4, and 7 EEHEE blueprints were used in design rounds 1-4 respectively. New blueprints were introduced in design round 3 that increased the protein length from 41 residues to 42 or 43 residues in order to increase the size of the potential hydrophobic core and increase the helix length (blueprints for design rounds 1-4 were 41, 41, 41/42/43, and 43 residues long respectively).

Each backbone structure produced above was used as the input to the Rosetta sequence design protocol FastDesign, also described previously (*33*). This protocol alternates between (a) a fixed-backbone Monte Carlo search in sequence and rotamer space, and (b) a fixed-sequence backbone relaxation step. This protocol begins with a softened repulsive potential and restores this potential to full strength across several cycles of design and relaxation. These design steps employ the Rosetta full-atom energy function. Design rounds and 1 and 2 employed the Talaris2013 version of the energy function (*40*); design round 3 employed the beta_july15 version of the energy function, and design round 4 employed the beta_nov15 version of the energy function (*19*, *55*). All design scripts are included in the data archive design_scripts.tar.gz; each script contains a complete protocol that includes both backbone construction and sequence design. The allowed amino acids at each position were restricted using the LayerDesign protocol (*34*); these restrictions are imposed separately from the design energy function for more efficient sampling and to account for design criteria not reflected in the energy function, such as solubility. In this protocol, positions on the designed structure are classified into “core”, "boundary", and "surface" layers according to their degree of burial, and polar amino acids are excluded from positions in the core layer while nonpolar amino acids are excluded from positions in the surface layer. Layer classification was performed using the "sidechain neighbors" protocol, which counts the number of neighboring residues in the region around the side chain of a given residue. Layer classification is performed on each structure individually and can change during the design process as the structure changes. The definitions of each layer (e.g. the level of burial required for a residue to be classified as core or boundary) were adjusted from design round to design round in order to increase the number of positions where hydrophobic amino acids were permitted. The specific layer definitions used at each stage are given in the included design scripts. For finer control over hydrophobic and polar positioning for the Round 4 designs, we manually specified the allowed amino acids at each position in the designed topologies using resfiles. Starting from design round 3, we included an amino acid composition-based energy term in the design energy function that penalized sequences possessing too few nonpolar amino acids. Finally, to limit the number of designs analyzed at the final selection stage, designs were filtered following sequence design using several basic criteria (primarily compactness and overall score; these filters are specified in the included design scripts).

After designing 2,000-40,000 finished designs per topology, we then analyzed and ranked these across diverse structural metrics inside and outside of Rosetta in order to select the final set of designs for experimental testing. In design rounds 1-3, this ranking was performed by selecting metrics of interest and assigning weights to each metric (again by hand) in order to produce a single composite design score, which was the sum of each metric multiplied by its weight. The selected metrics and their weights were adjusted until the top-ranking designs appeared optimal to the designer. Only a small number of metrics were used in design round 1 in order to ensure broad sampling of protein properties; additional metrics were added in further design cycles as new causes of failure were identified. Different weights were used for each different topology. All scoring metrics used are defined in the section *Methods: Definition of scoring metrics,* and the scores of all designs on all metrics are given in the attached data files in design_structural_metrics.tar.gz.

The metrics used for ranking designs in Round 1 evaluated each design's overall energy (total_score), β-sheet quality (hbond_lr_bb_per_res), packing (cavity_volume, degree, holes, AlaCount, pack), hydrophobic burial (buried_np, one_core_each, two_core_each, percent_core_SCN), agreement between sequence and local structure (mismatch_probability), solubility (exposed_hydrophobics), and hydrogen bond satisfaction (unsat_hbond). Based on the metrics that correlated with design success in Round 1 (as described in the text), we adjusted these weights to select new designs for Round 2, and also added additional metrics related to nonpolar burial (buried_minus_exposed, buried_over_exposed, contact_all) and the geometric similarity between 9-residue-long fragments of the designs and fragments of natural proteins of similar local sequence (avg_best_frag, worst6frags, worstfrag). We again adjusted these weights to select designs for Round 3, and added additional measures of local sequence-structure compatibility (abego_res_profile, p_aa_pp), fragment quality (avg_all_frags), packing (fa_atr_per_res, ss_sc), nonpolar burial (n_hphob_clusters, largest_hphob_cluster, hphob_sc_contacts), the spacing between nonpolar amino acids along the primary sequence (contig_not_hp_avg, contig_not_hp_max), and solubility (fxn_exposed_is_np). In design round 3, we also introduced restrictions on the sequence similarity of the selected βαββ designs (this topology featured the lowest amount of sequence variation between designs): rather than selecting the highest-ranking designs for testing, we selected designs in rank order while passing over designs that were more than 60% identical to a higher-ranking design already selected for testing (see Fig. S5 for sequence identity distributions of all designs).

For design round 4, we employed an automated design ranking scheme. All designs from all rounds were scored with ~50 structural metrics (see Table S3), and the structural metrics and experimental stability scores of the Round 1-3 designs were used to fit topology-specific linear regression, logistic regression, and gradient boosting regressions to predict experimental outcome (stability score) as a function of the design structural metrics. Models were fit using the scikit.learn package (*56*). Logistic regressions employed L1-regularization with C=0.1. Gradient boosting regressions employed 250 estimators with a tree depth of 5, the minimum samples per split set to 5, a learning rate of 0.01, and a least-squares loss function. All potential Round 4 designs were then ranked according to their predicted stabilities (linear and gradient boosting regression) or success probabilities (logistic regression), and designs were selected for testing in order of the lowest rank given to each design by any of the three regression models. In selecting designs for testing, we again passed over designs that were too similar to designs already selected for testing. We used a threshold of 70% identity for ααα and βαββ designs, and a threshold of 75% ββαββ designs (this was unnecessary for αββα designs). The median identity between designs sharing a topology remained 20-50% (Fig. S5).

DNA synthesis

All sequences were reverse translated and codon optimized using DNAworks2.0 (*57*). Sequences were optimized using *E. coli* codon frequencies despite being used for expression in yeast. Oligo libraries encoding designs and control sequences for design rounds 1 and 2 (12,472 sequences per round) were purchased from CustomArray, Inc. The oligo library for design round 3 (12,524 sequences) was purchased from Twist Bioscience. Oligo libraries for the point mutant library (13,564 sequences) and design round 4 (18,527 sequences, including the natural protein sequences) were ordered from Agilent Technologies in 27,000 feature format and selectively amplified out of the 27,000-sequence pool in the initial qPCR step. In order to amplify all sequences in a library as evenly as possible, we padded all sequences with extra residues until the amplified region of every oligo had the same length. The uniformity of all starting libraries is shown in Fig. S1. All libraries were 43 residues in length (not including 18bp adapter sequences on both ends), except design round 4 where all sequences were 50 residues in length. The saturation mutagenesis library also included 46-residue length oligos for the hYAP65 sequences. In the 43-residue libraries, the shorter EHEE and EEHEE sequences were padded with GSS or GS at the N-terminus. In design library 4, EHEE and EEHEE designed sequenced were again padded at the N-terminus with GSS or GS to reach 43 residues, and then all sequences were padded with N, G, and S residues at the C-terminus so that all sequences would be a uniform 50 residues in length.

DNA preparation and sequencing

Oligo libraries were amplified for yeast transformation in two qPCR steps. First, a 10 ng (CustomArray libraries) or 2.5 ng (Twist and Agilent libraries) quantity of synthetic DNA was amplified in a 25 µL reaction using Kapa HiFi Polymerase (Kapa Biosystems) for 10-20 cycles by qPCR. The number of cycles was chosen based on a test qPCR run in order to terminate the reaction at 50% maximum yield and avoid overamplification. Second, this reaction product was gel extracted to isolate the expected length product, and re-amplified by qPCR as before to generate sufficient DNA for high-efficiency yeast transformation. Each second qPCR reaction used 1/25th of the gel extraction product as template. This second PCR product was PCR purified and concentrated for transformation of EBY100 yeast using the protocol of (*58*) (3 µg of insert and 1.5 µg of cut vector per transformation). Yeast display employed a modified version of the pETcon vector (*59*) (known as “pETcon 3”), altered to remove a long single-nucleotide stretch near the cloning region. The amplified libraries included 40bp segments on either end to enable homologous recombination with the pETcon vector. Gel extraction and PCR purification were performed using QIAquick kits (Qiagen Inc).

DNA libraries for deep sequencing were prepared as above, except the first step started from yeast plasmid prepared from 5×10^7 to 1×10^8 cells by Zymoprep (Zymo Research). Cells were frozen at -80°C before and after the zymolase digestion step to promote efficient lysis. One-half the plasmid yield from the Zymoprep was used as the template for the first PCR amplification. Illumina adapters and 6-bp pool-specific barcodes were added in the second qPCR step. Unlike libraries prepared for transformation, DNA prepared for deep sequencing was gel extracted following the second amplification step. All libraries before and after selections were sequenced using Illumina NextSeq sequencing.

Yeast display proteolysis

S. cerevisiae (strain EBY100) cultures were grown and induced as in (*26*). Following induction, cell density ($O.D._{600}$) was measured by NanoDrop, and an amount of cells corresponding to 1 mL at O.D. 1 (12-15M cells) was added to each microcentrifuge tube for proteolysis. Cells were washed and resuspended in 250 µL buffer (20 mM NaPi 150 mM NaCl pH 7.4 (PBS) for trypsin reactions, or 20 mM Tris 100 mM NaCl pH 8.0 (TBS) for chymotrypsin reactions). Protolysis was initiated by adding 250 µL of room temperature protease in buffer (PBS or TBS) followed by vortexing and incubating the reaction at room temperature (proteolysis reactions took place at cell O.D. 2). After 5 minutes, the reaction was quenched by adding 1 mL of chilled buffer containing 1% BSA (referred to as PBSF or TBSF), and cells were immediately washed 4x in chilled PBSF or TBSF. Cells were then labeled with anti-c-Myc-FITC for 10 minutes, washed twice with chilled PBSF, and then sorted using a Sony SH800 flow cytometer using "Ultra Purity" settings. Events were initially gated by forward scattering area and back scattering area to collect the main yeast population, and then by forward scattering width and forward scattering height to separate individual and dividing cells (which were used for analysis) from cell clumps (which were discarded). Following these gates, cells were gated by fluorescence intensity in one-dimension (Fig. 1B), with the threshold separating displaying (fluorescent) from non-displaying (non-fluorescent) cells set at ~2,200 fluorescence units (Fig. 1B). Small adjustments were made to this gate based on daily conditions to maximize the separation between the major displaying and non-displaying populations. For each sort, we recorded the fraction of cells passing the fluorescence threshold before proteolysis (using cells from the same starting yeast population, but untreated with protease) and after proteolysis, and also recorded the total number of cells collected for each condition. These data are included in the file experiments.csv in the data archive counts_and_ec50s.tar.gz, and were used in the EC_{50} fitting procedure (see *Methods: EC_{50} estimation from sequencing counts*).

Design libraries 1-4 were were assayed at six protease concentrations over three sequential selection rounds. Trypsin assays used 0.07 μM, 0.21 μM, 0.64 μM, 1.93 μM, 5.78 μM, and 17.33 μM protease; chymotrypsin assays used 0.08 μM, 0.25 μM, 0.74 μM, 2.22 μM, 6.67 μM, and 20.00 μM protease. Selections using the lowest two concentrations of each protease (0.07 μM and 0.21 μM trypsin and 0.08 μM and 0.25 μM chymotrypsin) were performed starting from the naïve yeast library. The middle two selections (0.64 μM and 1.93 μM trypsin and 0.74 μM and 2.22 μM chymotrypsin) were performed starting from the post-selection 0.21 μM trypsin or 0.25 μM chymotrypsin cultures after 12-24 hours of growth and 12-24 hours of fresh induction. The highest concentration selections were performed starting from the post-selection 1.93 μM trypsin or 2.22 μM chymotrypsin cultures again following growth and re-induction.

The saturation mutagenesis library was assayed at six (trypsin) or eight (chymotrypsin) protease concentrations over four sequential selection rounds. Trypsin assays used 0.41 μM, 0.81 μM, 1.63 μM, 3.25 μM, 6.50 μM, and 13.00 μM protease; chymotrypsin assays used 0.21 μM, 0.42 μM, 0.84 μM, 1.69 μM, 3.38 μM, 6.75 μM, 13.50 μM, and 27.00 μM protease. As before, selections 1 and 2 were performed starting from the naïve library, selections 3 and 4 were performed starting from the selection 2 culture following growth and re-induction, selections 5 and 6 were performed starting from the selection 4 culture following growth and re-induction, and selections 7 and 8 (only done for chymotrypsin) were performed starting from the selection 6 culture following growth and re-induction. For trypsin, selection 6 was performed starting from the selection 5 culture following growth and re-induction.

Full results for each selection, including the fraction of displaying cells before and after each proteolysis experiment and the total number of cells collected at each stage, are given in the experiments.csv file in the data archive counts_and_ec50s.tar.gz.

Protease reagents

Trypsin-EDTA (0.25%) solution was purchased from Life Technologies and stored at stock concentration (2.5 mg/mL) at -20°C. α-Chymotrypsin from bovine pancreas was purchased from Sigma-Aldrich as lyophilized powder and stored at 1 mg/mL in TBS +100 mM $CaCl_2$ at -20°C. Each reaction used a freshly thawed aliquot of protease. The trypsin stock activity was measured to be 5,410 ± 312 BAEE units ($\Delta A_{253} \times 1{,}000 / 1$ minute) per mg in PBS buffer, pH 7.4, with 0.23 mM BAEE (Sigma-Aldrich). Using the Pierce Fluorescent Protease Assay Kit with the fluorescence protocol (ThermoFisher Scientific), 1 mg of the chymotrypsin stock was measured to have equivalent activity to 3.74 ± 0.31 mg of the trypsin stock in pH 7.4 PBS buffer at 25°C.

Processing of raw deep sequencing data

Each library in a sequencing run was identified via a unique 6 bp barcode. Following sequencing, reads were paired using the PEAR program (*60*). Reads were considered counts for a particular ordered sequence if the read (1) contained the complete NdeI cut site sequence immediately upstream from the ordered sequence, (2) contained the complete XhoI cut site sequence immediately downstream from the ordered sequence, and (3) matched the ordered sequence at the amino acid level (for sequences in designed libraries 1-4) or at the nucleotide level (sequences in the saturation mutagenesis library). A higher stringency was used for the saturation mutagenesis library due to the overall similarity of the sequences in the library.

EC$_{50}$ estimation from sequencing counts

To determine protease resistance from our raw sequencing data we built a probabilistic model of the cleavage and selection procedure and used this model to calculate maximum a posteriori estimates of the protease EC$_{50}$ of each member of the pool. To build the model, we assumed that proteolysis (i.e. any cleavage that results in detachment of the epitope tag) follows pseudo-first order kinetics, with a rate constant specific to each sequence. The fraction of surviving, tagged surface proteins for a given sequence after proteolysis is therefore:

$$f_{sprot} = e^{-k_p[E]t} \tag{1}$$

where k_p is a sequence-specific rate constant, $[E]$ is the concentration of protease and t is the reaction time.

In the assay, each cell has a labeling intensity proportional to the number of displayed proteins on its surface. Within the expressing population of cells, we assumed that the number of displayed proteins per cell is log-normally distributed, resulting in a distribution of labeling intensities $L_{cell} \propto \ln\mathcal{N}(\mu,\, \sigma^2)$ with sequence-independent location and scale parameters μ and σ. The fraction of cells collected at the labeling threshold $L_{cell} > L_s$ is then given by the cumulative distribution function:

$$\begin{aligned} Frac_{sel} &= 1 - CDF_{lognormal}\left(L_s, \mu, \sigma\right) && (2)\\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\Big[\frac{\ln L_s - \mu}{\sqrt{2}\sigma}\Big] && (3) \end{aligned}$$

Following proteolysis, labeling intensity is given by $L_{post} = L_{cell} \times f_{sprot}$ and cells are collected when $L_{post} > L_s$. With a fixed selection level L_s (defined as e^{c_s} in terms of log-intensity rather than absolute intensity) across selection rounds, the fraction of cells collected after proteolysis in each round is given by

$$\begin{aligned} Frac_{sel} &= 1 - CDF_{lognormal}\left(\frac{L_s}{f_{sprot}}, \mu, \sigma\right) && (4)\\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\Big[\frac{\ln \frac{L_s}{f_{sprot}} - \mu}{\sqrt{2}\sigma}\Big]\Big|L_s = e^{c_s} && (5)\\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\Big[\frac{c_s + k_p[E]t - \mu}{\sqrt{2}\sigma}\Big] && (6) \end{aligned}$$

For model fitting it is advantageous to describe protease stability in terms of a sequence-dependent variable EC_{50} and a sequence-independent variable K_{sel}. The EC_{50} for each sequence is defined as the protease concentration at which half of all cells displaying that sequence pass selection. K_{sel} is a constant term representing expression and collection conditions. Setting $Frac_{sel} = \frac{1}{2}$ allows us to define $k_p t$ in terms of the sequence-specific EC_{50}:

$$k_p t = \frac{\mu - c_s}{EC_{50}} \tag{7}$$

Substituting (7) into (6) and grouping the sequence-independent terms μ, σ, and c_s into K_{sel} yields $Frac_{sel}$ as a function of one sequence property (EC_{50}), one experimental variable for a given round ($[E]$), and the overall experimental conditions, which are assumed to be constant across all rounds (K_{sel}):

$$\begin{aligned} Frac_{sel} &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left[\frac{c_s + [E]\frac{\mu - c_s}{EC_{50}} - \mu}{\sqrt{2}\sigma}\right] && (8)\\ &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left[\frac{\mu - c_s}{\sqrt{2}\sigma}\left(\frac{[E]}{EC_{50}} - 1\right)\right]\Bigg| K_{sel} = \frac{\mu - c_s}{\sqrt{2}\sigma} && (9)\\ Frac_{sel}(EC_{50}, [E], K_{sel}) &= \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left[K_{sel}\left(\frac{[E]}{EC_{50}} - 1\right)\right] && (10) \end{aligned}$$

Finally, to account for non-specific selection (carryover) during sorting that can cause a sequence i to appear in the selected population even when $Frac_{sel,i} \approx 0$, we defined an adjusted $Frac_{sel}{}^{*}$ that includes the automatic propagation of a small amount a of the starting population into the selected population:

$$Frac_{sel}{}^{*}(EC_{50}, [E], K_{sel}, a) = a + (1 - a)Frac_{sel} \tag{11}$$

A fixed carryover amount of $a = 10^{-4}$ was used for all fitting.

We modeled each selection experiment as a set of discrete selection events producing both (A) a difference in the observed library population distribution after selection, and (B) a global selection rate during the sorting experiment. For each round of selection with enzyme concentration $[E]_{round}$, an observed input population distribution P_{in} is updated by a sequence-dependent proteolysis rate to produce an unobserved distribution of labeled cells P_{cleave}.

$$P_{cleave,i} = \frac{P_{in,i} \times Frac_{sel,i}{}^{*}(EC_{50,i}, K_{sel}, [E]_{round}, a)}{\sum_j P_{in,j} \times Frac_{sel,j}{}^{*}(EC_{50,j}, K_{sel}, [E]_{round}, a)} \tag{12}$$

for all sequences i, where the right-hand-side denominator of (12) normalizes that term so that $\sum P_{cleave} = 1$.

In each selection, n_{assay} total cells are examined, of which the cells n_{sel} passing the labeling threshold are collected. The n_{sel} collected cells are randomly selected from P_{cleave} during sorting to produced the observed post-selection distribution P_{sel}. Each library was analyzed by multiple rounds of selection, where the resulting population of a round may be used as the source population for a subsequent selection round at a higher protease concentration (see *Methods: Yeast display proteolysis* for details). For each round, the observed distribution of sequencing reads in the pre-selection library is used as P_{in}, which is normalized to sum to 1. The (non-normalized) post-selection distribution P_{sel} (which sums to the total number of cells collected n_{sel}) is computed by multiplying n_{sel}

by the observed normalized distribution of sequencing reads in the post-selection library. These definitions of P_{in} and P_{sel} assume that there are no sequence-dependent effects in amplification efficiency or sequencing efficiency. Only sequences in the designed libraries are included in P_{in} and P_{sel}; other sequences found in deep sequencing (due to errors in library synthesis, library amplification, or sequencing, or mutations during cell passaging) are not included in the model. In rare cases where a library sequence appears in P_{sel} despite being absent from the input distribution P_{in}, those observations in P_{sel} are ignored.

Because the model only considers sequences matching those in the designed library, the number of matching cells collected n_{sel} is smaller than the total number of cells collected as reported by the flow cytometer. To account for this, we crudely approximated n_{sel} as the number of cells collected by the flow cytometer multiplied by the fraction of sequencing reads that matched sequences in the designed library (i.e. if 200,000 cells were collected in a sort and 75% of sequencing reads for that sort matched library sequences, we assumed in the model that 150,000 cells containing library sequences had been collected). In rare cases where the total number of matching sequencing reads for a given library was smaller than the estimated n_{sel} (thus the statistical error was limited by sequencing rather than sorting), the number of matching sequencing reads was used as n_{sel}. Finally, to calculate the total number of cells observed (containing library sequences) n_{assay}, we assumed that the overall collection rate (i.e. one collected cell for every twenty observed displaying cells) could be used as a proxy measure of the collection rate for library sequences (as would be true if library sequences dominated the displaying cell population, or if the collection rates for displaying cells with library and non-library sequences were approximately equal). We calculated the overall collection rate directly from the flow cytometry data as the fraction of initially displaying cells that remained displaying following proteolysis, and then calculated n_{assay} by dividing n_{sel} by the overall collection rate.

The complete model log-likelihood is the sum of the data-log likelihoods of P_{sel} and n_{sel} and prior likelihoods over the fit parameter EC_{50}, taking P_{in} and n_{assay} as given. K_{sel} was initially treated as a fit parameter as well, but for consistency between all libraries, we fixed K_{sel} at 0.8 for all analysis in this work. The log-likelihood of the observed population P_{sel} was modeled as a multinomial distribution of n_{sel} independent selections from P_{cleave}:

$$Mn(P_{sel} \mid n_{sel}, P_{cleave}) \tag{13}$$

The log-likelihood of the observed global selection rate was modeled as a binomial distribution of selection events, where the overall selection probability $Frac_{sel,pop}{}^*$ is the weighted average of each $Frac_{sel}{}^*$:

$$Bn(n_{sel} \mid n_{assay}, Frac_{sel,pop}{}^*) \mid Frac_{sel,pop}{}^* = \sum_i Frac_{sel,i}{}^* P_{in,i} \tag{14}$$

for all sequences i. Uniform priors covering the range of experimentally relevant values were used for the model parameters. The MAP estimate of the model parameters is found by optimizing the expression:

$$\underset{EC_{50}}{\operatorname{argmax}} \sum_{r \in round} Mn_r + Bn_r \tag{15}$$

The 95% credible intervals (defined as the central 95% of the probability density) for all EC$_{50}$s were also estimated from the likelihood expression given in (15). The actual number of sequencing counts, as well as the number of counts predicted for all sequences at all selection stages according to the fitted model parameters, are given in the supplementary data archive counts_and_ec50s.tar.gz. All model components were implemented in Python via PyMC3 (*61*) and Theano (*62*). The complete model fitting code is available at
https://github.com/asford/protease_experimental_analysis.

For the analysis of design success rates and design features correlating with success in Figs. 2, S7, and S8, we excluded sequences whose model-estimated EC$_{50}$ credible intervals were large. To include as much data as possible, we used a permissive threshold: designs were included in the analysis if their chymotrypsin and trypsin stability score 95% credible intervals (directly taken from the EC$_{50}$ credible intervals) were smaller than 0.95 stability score units (A factor of 9 in [protease]; the equivalent of two rounds of sorting). These thresholds excluded from analysis 14%, 30%, 0.7%, and 1% of sequences from design rounds 1-4 respectively. Credible intervals were much narrower in the later rounds due to improved DNA libraries and better representation of each design in sorting; see Fig. S1 and Fig. S2. Despite the permissive thresholds, the median 95% credible interval width for stability scores for sequences included in the analysis was 0.14 stability score units, and 95% of the credible intervals were smaller than 0.48 stability score units.

Unfolded state model

We trained a model for the expected protease EC$_{50}$ of an unfolded protein using our stability data on scrambled sequences. We used both fully scrambled sequences and hydrophobic-polar pattern-preserving scrambled sequences as training data ($\sim$18,000 sequences in total, see Fig. S3). Only sequences with EC$_{50}$ value 95% credible intervals smaller than a factor of 3 in [protease] were used for model fitting (protease concentrations increased by this amount at each selection step).

To define the model, we separated the cutting rate into a fixed rate for the constant regions of the fusion construct and an individual rate of cutting at each site i in the inserted sequence. This was done by rearranging equation (7):

$$EC_{50} = \frac{\mu - c}{k_f t + \sum_{i=1}^{n} k_i t} \tag{16}$$

where k_f is the pseudo-first order rate constant for the constant regions of the fusion construct in ${M_{enzyme}}^{-1}\ s^{-1}$, k_i is the cleavage rate after amino acid i, n is the number of residues over which cleavage is considered (the residues in the inserted sequence as well as flanking residues whose cleavage rates may be modified by the presence of the inserted sequence), and t is time. If we assume that (1) the inserted sequence cannot affect the cleavage rate of the constant sequence, and (2) that the inserted sequence is completely uncleaved (all k_i = 0), then the EC_{50} reaches a maximum that is independent of the inserted sequence:

$$EC_{50\,\mathrm{max}} = \frac{\mu - c}{k_f t} \tag{17}$$

By dividing the numerator and denominator on the right-hand-side of (16) by $k_f t$, we can re-write (16) as:

$$EC_{50} = \frac{EC_{50\,\mathrm{max}}}{1 + \sum_{i=1}^{n} \frac{k_i}{k_f}} \tag{18}$$

We modeled k_i/k_f as a function of the 9-residue-long local sequence surrounding sequence position i. In other words, the cut rate at site i in the model depends on the amino acid identities at sites $i - 4$ through $i + 4$, referred to as sites P5 to P4' in protease nomenclature. The effects of the sequence at these positions is implemented through a position-specific scoring matrix (PSSM) with coefficients for all 19 amino acids (excluding cysteine) at positions P5-P4' around a potential cut site. The model for k_i/k_f as a function of the local sequence is given below:

$$\frac{k_i}{k_f} = \frac{k_{\mathrm{max}}}{1 + \exp\left(c_0 - \sum_{site=P5}^{P4} PSSM(aa_{site}, site)\right)} \tag{19}$$

where aa_{site} is the amino acid identity at $site$. The parameters of the full model are $EC_{50\,\mathrm{max}}$, k_{max}, c_0, and the $19 \times 9 = 171$ elements of the PSSM. Distributing the c_0 term into the PSSM coefficients would not affect the model (c_0 adds no additional model freedom), but including the term aided model fitting. Positive PSSM coefficients lead to a smaller denominator in eqn. (19), an increased cutting rate k_i, and a lower predicted EC_{50} by eqn. (18). The model parameters (referred to collectively as θ) were trained by minimizing the logarithmic error between the model predicted EC$_{50}$s and the observed EC$_{50}$s over the training set of scrambled sequences. We used a combination of squared-error and absolute error in the objective function to provide slightly more tolerance for large outliers than squared-error alone.

$$\underset{\theta}{\mathrm{argmin}} \sum_{seq} \left(\log(EC_{50,obs}) - \log(EC_{50}(seq,\theta))\right)^2 + 0.25 \times \mathrm{abs}(\log(EC_{50,obs}) - \log(EC_{50}(seq,\theta))) \tag{20}$$

We trained the model starting from a uniform PSSM by iterating between fitting only the P1 component of the PSSM, all other positions of the PSSM, and the $EC_{50\,\mathrm{max}}$, k_{max}, and c_0 terms. The model fitting code, implemented in Python using Theano (62), is provided at https://github.com/asford/protease_experimental_analysis. The final PSSM elements and the overall agreement to the data are shown in Fig. S3. We validated the model using three-fold cross-validation (three separate models built by excluding a different one-third of the data at a time, followed by predicting each EC_{50} using the model that did not encounter that sequence during training). The cross-validated root-mean-squared errors (RMSE) for trypsin and chymotrypsin are 2% (trypsin) and 5% (chymotrypsin) higher than the RMSEs of the models trained using all data without cross-validation, indicating minimal overfitting. The predictions made by the cross-validated models are very similar to the predictions made by the models trained on the complete dataset (Fig. S4 C,D).

All stability scores reported in the manuscript and used to generate the figures were calculated using the final version of the unfolded state model, trained using the scrambled sequence data from all four design rounds. Obviously, data on the full set of scrambled sequences was not available at earlier stages of the work. The data analysis after each design round employed earlier versions of the unfolded state model that were trained using the scrambled sequence data that had been collected up to that point. However, the final model predictions used for the manuscript are very similar to the model predictions made when only the data from Round 1 are used for training (Fig. S4, A,B).

Because the unfolded state model is trained on EC_{50}s of scrambled sequences and not on designed sequences, a systematic bias may be introduced that would cause scrambled sequences to receive lower stability scores than designed sequences (the stability score is the deviation of each sequence's measured EC_{50} from the unfolded state models predicted EC_{50}; if the model were overfit, the sequences used in training would have incorrectly low deviations). However, the cross-validation results in Fig. S4 C and D indicate that only minimal overfitting is present in the model parameters. To further quantify possible bias in the model parameters, we examined the distribution of predicted unfolded state EC_{50} values for scrambled sequences (which were used in training) and designed sequences (which were not). We would expect these distributions to be the same because the designed and scrambled sequences are very similar at the sequence level. However, on average, the predicted unfolded state EC_{50} values for the scrambled sequences are higher than the predicted EC_{50} values for the designed sequences, which biases the scrambled sequences to appear less stable, although this effect is small (Fig. S4 E,F). This bias likely results from overfitting of EC_{50} values for partially folded scrambled sequences. Overall, the small bias (0.15-0.16 units of stability score on average) between designed sequences and scrambled sequences does not change the conclusion of Fig. 2 that hundreds of the designs at each stage are many times more stable than the scrambled sequences, often by 0.5-1.0 stability score units or more.

Protein expression and purification

Two different expression vectors were used to purify the designs chosen for biophysical analysis; Table S1 lists the expression approach used for each design. Most designs were expressed as isolated domains with an additional 21-residue N-terminal sequence containing a His-tag and thrombin cleavage site to aid purification (full sequence: MGSSHHHHHHSSGLVPRGSHM). Genes encoding these designs were obtained from GenScript in the pET-28b+ expression vector. The remaining designs were expressed as fusions with the yeast SUMO domain Smt3 using the custom vector pCDB24. Genes encoding these designs were obtained as gBlocks from IDT and inserted into the pCDB24 vector via Gibson assembly (*63*).

All designs were expressed in E. coli BL21* (DE3) cells (Invitrogen). Starter cultures were grown overnight at 37°C in Luria-Bertani (LB) medium overnight with added antibiotic (50 µg/ml carbenicillin for SUMO expression or 30 µg/ml kanamycin for pET-28b+ expression). These overnight cultures were used to inoculate 500 mL of Studier autoinduction media (*64*) supplemented with antibiotic, and grown overnight. Cells were harvested by centrifugation at 4°C, resuspended in 25 mL lysis buffer (20 mM imidazole in PBS containing DNAse and protease inhibitors), and lysed by sonication or by microfluidizer. PBS buffer contained 20mM $NaPO_4$, 150mM NaCl, pH 7.4. After removal of insoluble material, the lysates were loaded onto nickel affinity gravity columns to purify the designed proteins by immobilized metal-affinity chromatography (IMAC). Designs expressed as fusions to the SUMO domain were then cleaved from the domain using the Yeast SUMO protease Ulp1 and dialyzed overnight in PBS at 4°C to remove excess imidazole before a second IMAC step was used to remove the SUMO tag following cleavage.

For expression of ^{13}C-^{15}N-labeled protein for NMR analysis, the plasmids were transformed into the Lemo21 E. coli expression strain (NEB) and plated on M9/glucose plates containing kanamycin to 50 ug/mL and chloramphenicol to 34 ug/mL, grown at 37°C overnight. For the starter culture, a single colony was inoculated into a 250mL baffled flask containing 50mL of Luria-Bertani medium, with kanamycin to 50 ug/mL, chloramphenicol to 34 ug/mL, and grown for approximately 18 hours at 37°C, shaking at 225rpm. 10 mL of the starter culture was then transferred to a 2L baffled flask containing 0.5L of Terrific Broth (Difco), with 25mM Na_2HPO_4, 25mM KH_2PO_4, 50mM NH_4Cl, 5 mM Na_2SO_4, kanamycin to 50 ug/mL, and chloramphenicol to 34 ug/mL. This expression culture was grown at 37°C to an OD_{600} of approximately 1.0, then removed from the flask and spun at 4000rpm for 15 minutes to pellet the cells. The Terrific Broth was removed, and the cells were washed briefly with 30 mL of PBS. The cells were then transferred to a new 2L baffled flask containing 0.5 L of labeled media (25mM Na_2HPO_4, 25mM KH_2PO_4, 50mM $^{15}NH_4Cl$, 5 mM Na_2SO_4, 0.2% (w/v) ^{13}C glucose), kanamycin to 50 ug/mL and chloramphenicol to 34 ug/mL. The cells were allowed to recover at 37°C for 30 minutes, then IPTG (Carbosynth) was added to 1mM and the temperature was reduced to 20°C. The cells were harvested the following day and purified by IMAC. The labeled NH_4Cl and glucose were obtained from Cambridge Isotopes.

Size-exclusion chromatography

Following IMAC, designs (labeled and unlabeled) were further purified by size-exclusion chromatography on ÄKTAxpress (GE Healthcare) using a Superdex 75 10/300 GL column (GE Healthcare) in PBS buffer. The monomeric fraction of each run (typically eluting at the 15 mL mark) was collected and immediately analyzed by CD or flash frozen in liquid N_2 for later analysis.

Circular dichroism

Far-ultraviolet CD measurements were carried out with an AVIV spectrometer, model 420. Wavelength scans were measured from 260 to 195 nm at 25 and 95°C. Temperature melts monitored dichroism signal at 220 nm in steps of 2°C/minute with 30s of equilibration time. Wavelength scans and temperature melts were performed using 0.35 mg/ml protein in PBS buffer (20mM $NaPO_4$, 150mM NaCl, pH 7.4) with a 1 mm path-length cuvette. Chemical denaturation experiments with guanidinium hydrochloride (GuHCl) were performed using an automatic titrator with a protein concentration of 0.035 mg/ml and a 1 cm path-length cuvette with stir bar. The GuHCl concentration was determined by refractive index in PBS buffer. The denaturation process monitored dichroism signal at 220 nm in steps of 0.2 M GdmCl with 1 minute mixing time for each step, at 25°C. Protein concentrations were determined by absorbance at 280 nm measured using a NanoDrop spectrophotometer (Thermo Scientific) using predicted extinction coefficients (*65*). Protein concentrations for designs lacking aromatic amino acids were measured by Qubit protein assay (ThermoFisher Scientific).

Melting temperatures were determined by first smoothing the data with a Savitsky-Golay filter of order 3, then approximating the smoothed data with a cubic spline to compute derivatives. The reported Tm is the inflection point of the melting curve. Chemical denaturation curves were fitted by nonlinear regression to a two-state unfolding model with six-parameters: the folding free energy, m-value, and linear pre- and post-transition baselines with individual slope and intercepts (*66*).

NMR structure determination

NMR data acquisition was carried out at 25°C (HHH_rd1_0142, EHEE_rd1_0284, and EEHEE_rd3_1049) or 15°C (HEEH_rd4_0097) on Bruker spectrometers operating at 600 or 800 MHz, and equipped with cryogenic probes. All 3D spectra were acquired with non-uniform sampling schemes in the indirect dimensions and were reconstructed by multi-dimensional decomposition software MDDNMR (*67*) or (*68*), interfaced with NMRPipe (*69*). Conventional backbone and NOESY spectra were acquired as described previously (*70*), and the automated program ABACUS (*71*) was used to aide in the assignment of backbone and sidechain resonances. Initial automated NOE assignments and structure calculations were performed using the noeassign module in CYANA 3.0 (*72*). The best 20 of 100 CYANA structures from the final cycle were refined with CNSSOLVE (*73*) by performing a short restrained molecular dynamics simulation in explicit solvent (*74*). The NMR structures of the constructs are comprised of the final 20 refined structures.

Fragment analysis

To evaluate agreement between sequence and structure for a given designed protein, we used Rosetta's standard fragment generation protocol (*17*) to select 200 fragments (9-residue length segments) from natural protein crystal structures for each 9-residue-long segment of the designed protein. The fragments were chosen so that their sequence and secondary structure were as similar as possible to the sequence and *predicted* secondary structure of the designed protein segment (predicted using PSIPRED (*75*)). If these fragments are highly geometrically similar to the designed segment (measured by RMSD), this indicates that the designed sequence preferentially adopts the designed fold even at the local level, because each local sequence segment is commonly found in its designed local structure when found in solved protein structures. In Fig. 2C, geometric similarity was quantified as the average RMSD of all 200 fragments at all positions (the "avg_all_frags" metric described in *Methods: Definition of scoring metrics*). Other measures of agreement are also described in that section.

Mutational stability effects

Instead of using the minimum of the trypsin and chymotrypsin stability scores as an overall stability score for sequences in the point mutant library (as we did in Figs. 2 and 5), we took advantage of the hundreds of mutants available for each protein to calibrate the trypsin and chymotrypsin stability scales in relation to each other for each set of mutants (i.e. mutants of the same wild-type protein). For example, mutations in EHEE_rd1_0882 that cause a chymotrypsin stability score change of 1.0 typically cause a trypsin stability score change of 1.2 (i.e. the slope of the best-fit line is 1.2; the r^2 for the two datasets is 0.77). However, mutations to EEHEE_rd3_0037 that cause a chymotrypsin stability score change of 1.0 cause a much larger trypsin stability score change of 2.6 ($r^2 = 0.71$). Because each set of mutants had a characteristic slope, we used these slopes to combine the trypsin and chymotrypsin measurements and compute a consensus stability score for each mutant. These consensus stability scores were assigned in four steps. First, we identified the subset of mutants for each wild-type protein with high-confidence EC_{50} values (i.e. those that were precisely measured and were within the dynamic range of protease concentrations tested) (Fig. S9A). Second, these high-confidence measurements were then used to determine the slope and intercept of the best-fit line between the trypsin and chymotrypsin stability scores for each protein by orthogonal distance regression (Fig. S9B). Third, we mapped each mutant of a given protein onto the best-fit line for that protein at one of three positions: the nearest point on the fit line, the point on the fit line at an identical x-coordinate (chymotrypsin stability score), or the point on the fit line at an identical y-coordinate (trypsin stability score) (Fig. S9B). See Fig. 9B for examples of how this mapping was performed in order to maximize the effective dynamic range of the consensus stability score. Finally, after mapping all points onto the fit line, we used the x-coordinate of the mapped point (the location of that point on the chymotrypsin axis) as the overall consensus stability score for each mutant. All EC_{50}s and stability scores are provided in the supplementary data archive stability_scores.tar.gz.

In examining the best-fit lines between the trypsin and chymotrypsin measurements for each mutant set, we observed that mutants whose predicted unfolded state chymotrypsin EC_{50} values changed significantly from the predicted unfolded state wild-type EC_{50} value were often outliers in the fit. These outliers suggested that the chymotrypsin unfolded state model was oversensitive to the effects of single amino acid changes, distorting the fits. To improve the estimation of consensus stability scores, we restricted the deviation between each mutant's predicted unfolded state EC_{50} and the wild-type predicted unfolded state EC_{50} to a factor of 2.64 in [chymotrypsin] ($2^{1.4}$). This only affected 1.2% of mutants, and would not have affected any results in Fig. 1 (no mutants in Fig. 1 deviate by this amount from the wild-type predicted unfolded state EC_{50} value.)

The trypsin, chymotrypsin, and consensus stability scores for all mutants are shown in Fig. S10. The average stability effects of each amino acid shown in Fig. 4E-L were calculated using the consensus stability scores described above. To compute the average stability effects using the data, we used the average stability score of the A, E, H, I, M, T, and V mutants at each position as the "baseline" stability at each position (i.e. the average stability score of these mutants was used as the zero-point for a new position-specific stability scale). These amino acids were chosen because they included the different types of amino acid physical properties (polar and hydrophobic, large and small) and because these amino acids generally have minimal impact on a sequence's unfolded state predicted EC_{50} with either trypsin or chymotrypsin. We then computed the stability of each amino acid at each position relative to this baseline. Finally, we averaged these re-zeroed stability scores across all the different protein sites in a given category (i.e. polar helical positions, edge strand positions, etc.) to determine the average stability effect of each amino acid for that category. The re-zeroing procedure for each site did not affect the relative average stabilities of the different amino acids shown in the figure, but it did lower the associated standard error by removing irrelevant variation in overall protein stability from the measurement. To depict the data in Fig. 4E-L, the average stability effects were adjusted a final time to set the mean value of all 20 amino acids to zero. For Fig. 4E-H, helical positions were considered to be polar if their wild-type (designed) amino acid was D, E, H, K, N, Q, R, S, T, or Y. The first and last helical turns were defined as the first and last three residues of each helix.

These stability effects were computed using the mutational data on all designed proteins in Fig. S10 except EEHEE_rd3_1702 and HEEH_rd3_0726. EEHEE_rd3_1702 was excluded from this analysis because its mutational profile was inconsistent with its designed structure (see Fig.

S10M). HEEH_rd3_0726 was excluded from this analysis because the chymotrypsin unfolded state background model appeared to perform poorly on this protein (many mutants appear stabilizing in exactly the amount by which they change the expected background proteolysis rate, i.e. with no change in EC_{50}).

Natural protein compilation

Our 1,178 natural proteins were compiled by querying the PDB on February 4th, 2016 for all structures containing only protein (no lipid, carbohydrate, or nucleic acid), with 1 chain per asymmetric unit, chain length 20-50aa, and no modified residues. We then manually filtered this list to remove all sequences containing Cys residues. Pfam sequences were collected from the seed database of Pfam 28.0, taking the first representative of all families lacking a Cys residue.

Conservation analysis in naturally occurring proteins

Homologous sequences for villin HP35, pin1 WW-domain, and hYAP65 WW-domain were identified using HHblits (*76*) with an e-value cutoff of 1e-10 and 4 iterations. HHfilter was used to remove sequences that were more than 90% identical or that covered less than 90% of the query sequence (*77*). Bits of conservation were calculated using WebLogo 3.3 (*78*).

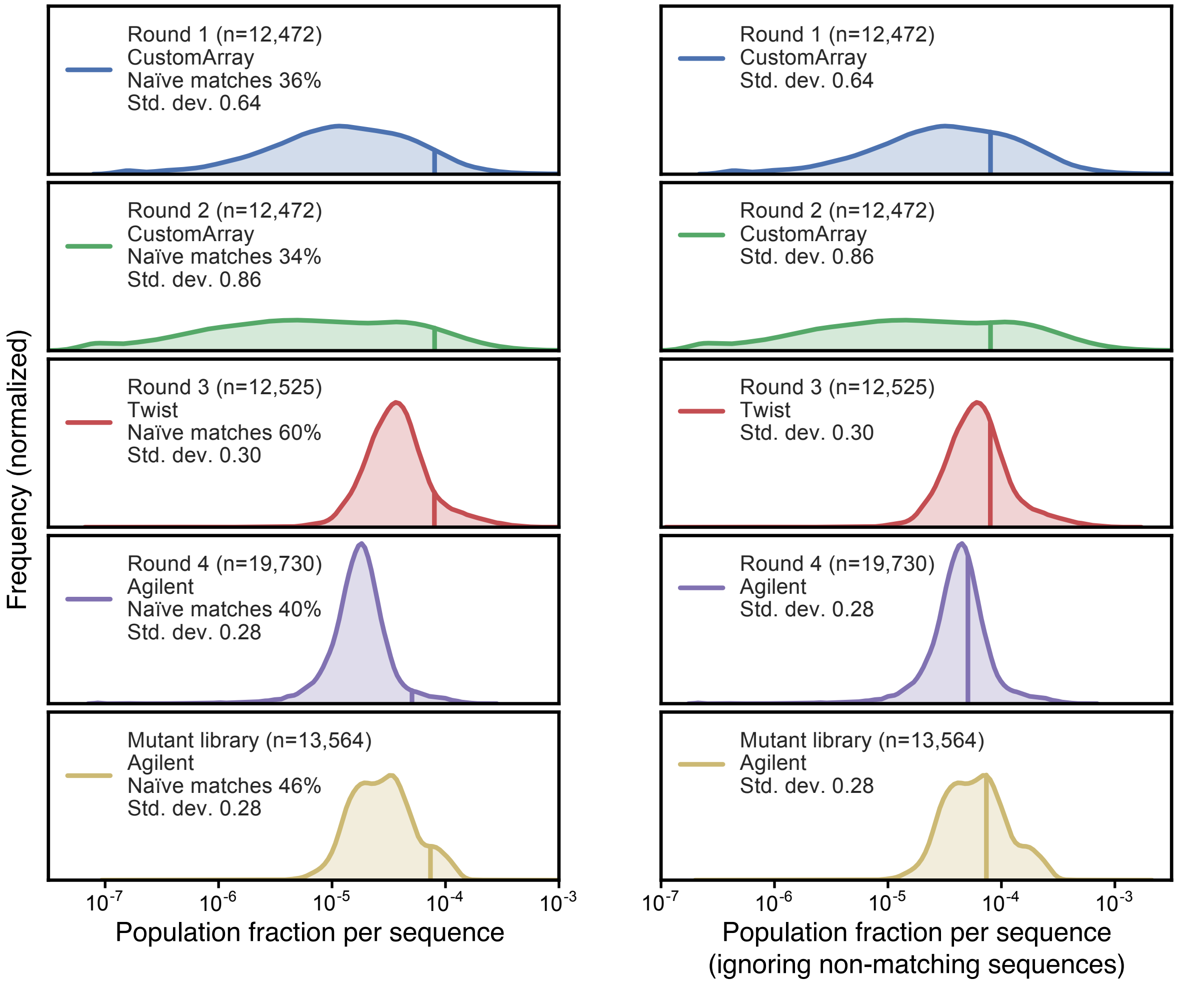


Fig. S1. Oligo library synthesis (OLS) accuracy and uniformity of libraries before selection. The population fraction for sequence *i* is defined as the number of deep sequencing reads for sequence *i* divided by either the total number of reads (left column) or the total number of reads that match any sequence specified for synthesis (i.e. excluding non-matching reads caused by sequencing errors, amplification errors, OLS errors, etc.) (right column). Each plot shows the distribution of these population fractions for all *n* specified sequences as a kernel density estimate. The colored vertical line inside each distribution shows the expected population fraction if the library were perfectly uniform with no non-matching sequences; i.e. $\log_{10}(1/n)$. For each of the five libraries used, we list the number of specified sequences in the library (*n*), the library manufacturer (CustomArray, Twist, or Agilent), the percentage of reads in the unselected (naïve) library that matched the specified sequences (amino acid identity-level matches for Rounds 1-4 and nucleotide identity matches for the mutant library), and the standard deviation of $\log_{10}$ population fractions for all sequences. Unselected libraries were sequenced following PCR amplification, high-efficiency yeast transformation and growth, and yeast plasmid preparation as described in *Methods: DNA preparation and sequencing*; these steps may add nonuniformity beyond that present in the manufactured library.

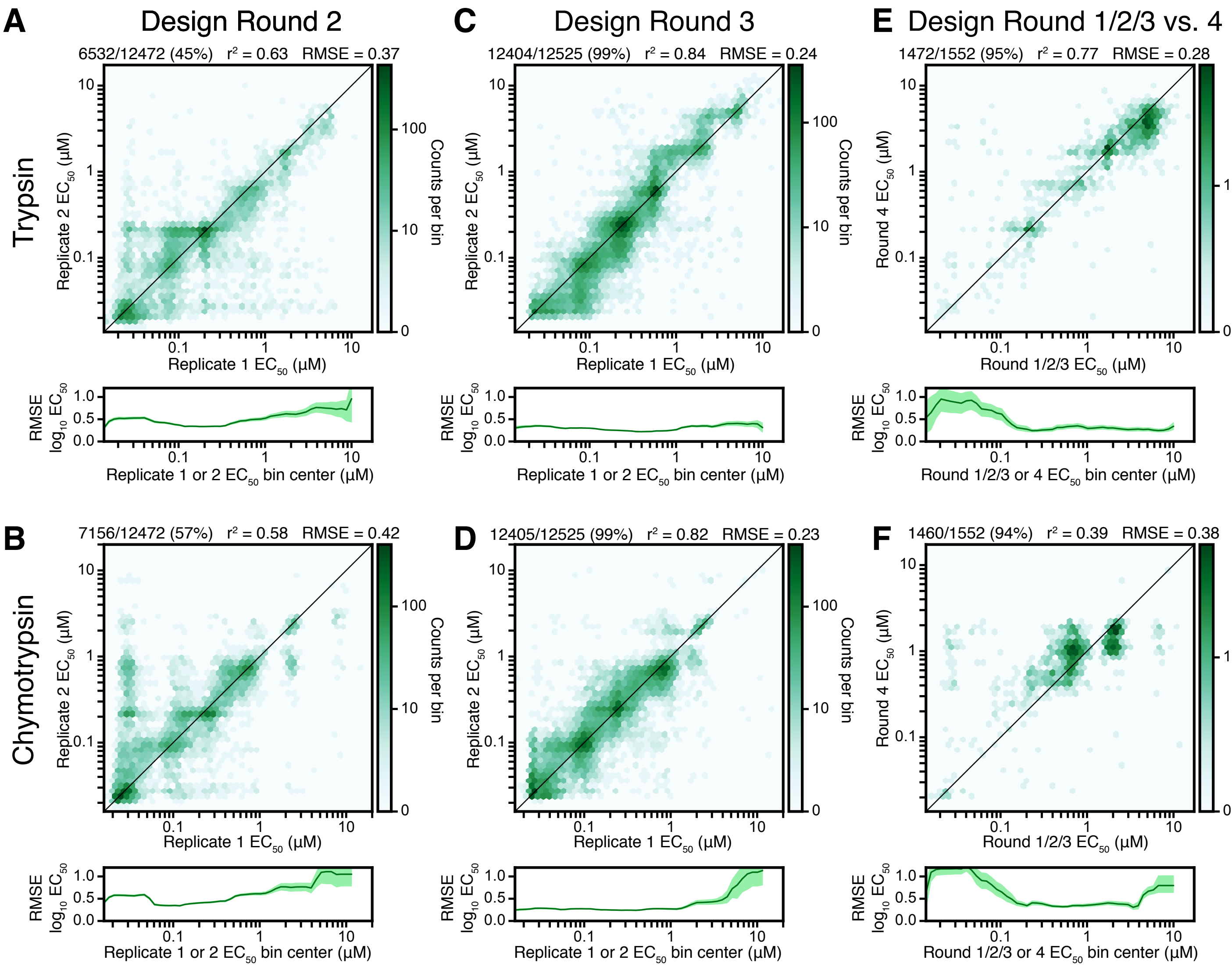
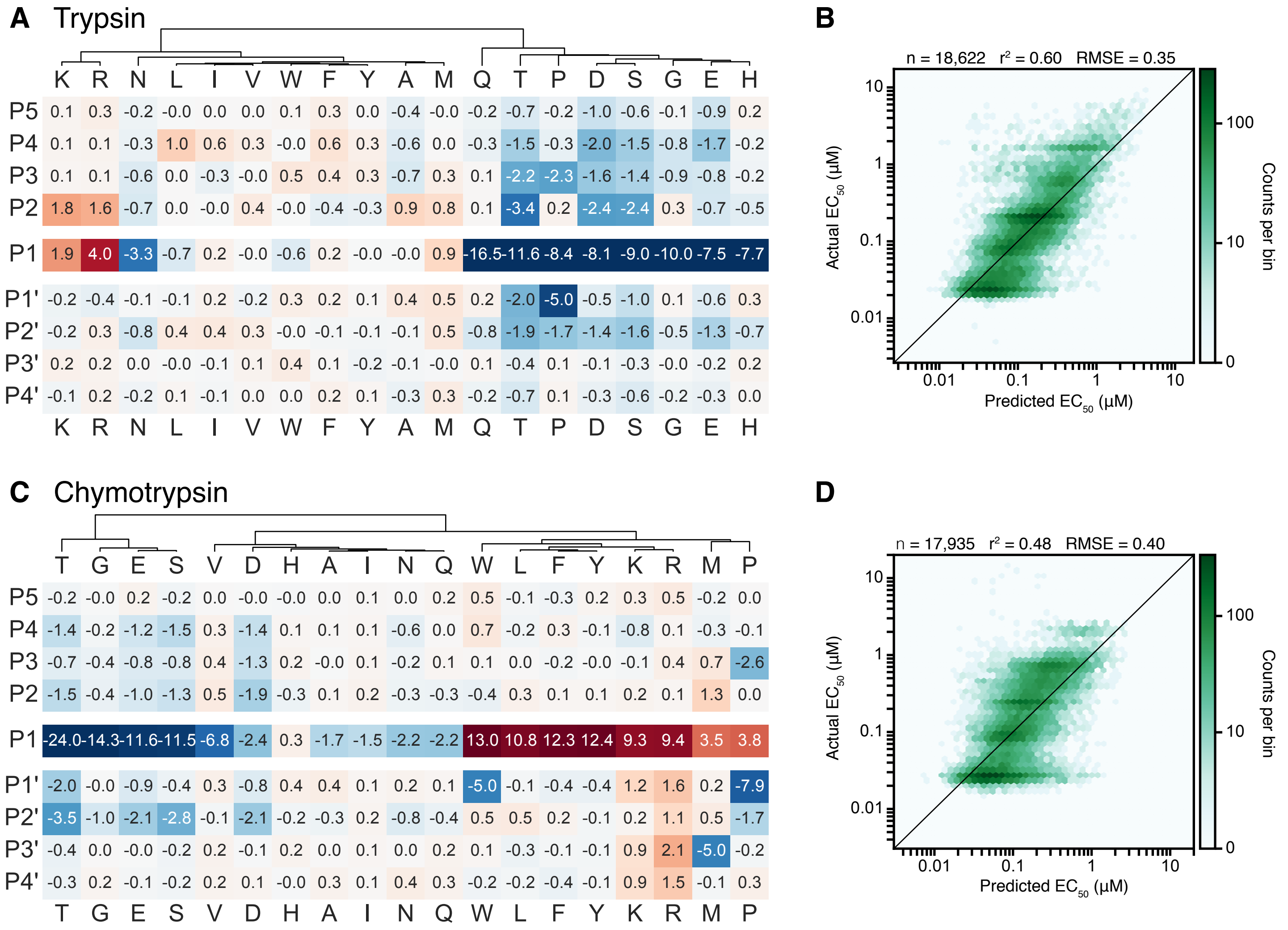


Fig. S2. Reproducibility of EC$_{50}$ measurements within and across libraries. To examine reproducibility of raw EC$_{50}$ values when a particular library is measured twice, we performed complete biological replicates of yeast display proteolysis measurements for the Round 2 library (**A, B**) and the Round 3 library (**C, D**) with both trypsin (**A, C**) and chymotrypsin (**B, D**). Replicates started from the same transformed yeast culture but were independent from the first proteolysis step onward. Agreement between replicates is shown as a 2D histogram (upper plot) counting the number of sequences with a given EC$_{50}$ in each replicate, colored on a log scale. The black diagonal line indicates perfect agreement between replicates and is not a best-fit line. Only sequences passing our confidence thresholds in both replicates are used in this analysis (see *Methods, EC$_{50}$ estimation from deep sequencing*. Note that credible intervals are computed per-replicate based on sequencing counts; consistency between replicates is then evaluated *after* these credible intervals have been computed). Each plot is annotated at the top with the total number and percentage of sequences passing the confidence thresholds, the overall r^2 between the replicates, and the root-mean-squared error (RMSE) between the log$_{10}$ EC$_{50}$ values of the two replicates. Below each 2D histogram, we plot a moving RMSE calculated across all regions of the EC$_{50}$ dynamic range using a window of size of 0.5 (in log$_{10}$ EC$_{50}$ units); a pair of EC$_{50}$ values from

replicates 1 and 2 is included in the RMSE calculation for a given window if either EC_{50} value falls within that window. The 95% confidence interval of the RMSE is shown in light green shading, computed by bootstrapping. Note that a large fraction (~50%) of EC_{50} measurements from Round 2 (**A, B**) did not pass our confidence thresholds for one or both replicates. This is due to poor uniformity in the starting Round 2 library, causing many sequences to be undersampled in sorting (see Fig. S1). The fraction of sequences passing confidence thresholds is greatly improved in the more-uniform Round 3 library (99% passing), and the RMSE between replicates improves as well (**C, D**). To examine reproducibility of raw EC_{50} values when a given sequence is assayed in a new library context, 1552 sequences with high trypsin or chymotrypsin stability were included in the Round 4 library; agreement between initial and re-measured EC_{50} values is plotted and summarized as before for trypsin (**E**) and chymotrypsin (**F**). Note that the re-measured sequences were biased to have high chymotrypsin stability, limiting the dynamic range and leading to a poor correlation between original and re-measured chymotrypsin EC_{50} values (r^2 0.39). A better correlation is seen with trypsin (r^2 0.77) across a more complete EC_{50} dynamic range; this is comparable to the correlation seen between replicates of the same library.



	K	R	N	L	I	V	W	F	Y	A	M	Q	T	P	D	S	G	E	H
P5	0.1	0.3	-0.2	-0.0	0.0	0.0	0.1	0.3	0.0	-0.4	-0.0	-0.2	-0.7	-0.2	-1.0	-0.6	-0.1	-0.9	0.2
P4	0.1	0.1	-0.3	1.0	0.6	0.3	-0.0	0.6	0.3	-0.6	0.0	-0.3	-1.5	-0.3	-2.0	-1.5	-0.8	-1.7	-0.2
P3	0.1	0.1	-0.6	0.0	-0.3	-0.0	0.5	0.4	0.3	-0.7	0.3	0.1	-2.2	-2.3	-1.6	-1.4	-0.9	-0.8	-0.2
P2	1.8	1.6	-0.7	0.0	-0.0	0.4	-0.0	-0.4	-0.3	0.9	0.8	0.1	-3.4	0.2	-2.4	-2.4	0.3	-0.7	-0.5
P1	1.9	4.0	-3.3	-0.7	0.2	-0.0	-0.6	0.2	-0.0	-0.0	0.9	-16.5	-11.6	-8.4	-8.1	-9.0	-10.0	-7.5	-7.7
P1'	-0.2	-0.4	-0.1	-0.1	0.2	-0.2	0.3	0.2	0.1	0.4	0.5	0.2	-2.0	-5.0	-0.5	-1.0	0.1	-0.6	0.3
P2'	-0.2	0.3	-0.8	0.4	0.4	0.3	-0.0	-0.1	-0.1	-0.1	0.5	-0.8	-1.9	-1.7	-1.4	-1.6	-0.5	-1.3	-0.7
P3'	0.2	0.2	0.0	-0.0	-0.1	0.1	0.4	0.1	-0.2	-0.1	-0.0	0.1	-0.4	0.1	-0.1	-0.3	-0.0	-0.2	0.2
P4'	-0.1	0.2	-0.2	0.1	-0.1	0.0	0.0	0.2	0.1	-0.3	0.3	-0.2	-0.7	0.1	-0.3	-0.6	-0.2	-0.3	0.0
	K	R	N	L	I	V	W	F	Y	A	M	Q	T	P	D	S	G	E	H

	T	G	E	S	V	D	H	A	I	N	Q	W	L	F	Y	K	R	M	P
P5	-0.2	-0.0	0.2	-0.2	0.0	-0.0	-0.0	0.0	0.1	0.0	0.2	0.5	-0.1	-0.3	0.2	0.3	0.5	-0.2	0.0
P4	-1.4	-0.2	-1.2	-1.5	0.3	-1.4	0.1	0.1	0.1	-0.6	0.0	0.7	-0.2	0.3	-0.1	-0.8	0.1	-0.3	-0.1
P3	-0.7	-0.4	-0.8	-0.8	0.4	-1.3	0.2	-0.0	0.1	-0.2	0.1	0.1	0.0	-0.2	-0.0	-0.1	0.4	0.7	-2.6
P2	-1.5	-0.4	-1.0	-1.3	0.5	-1.9	-0.3	0.1	0.2	-0.3	-0.3	-0.4	0.3	0.1	0.1	0.2	0.1	1.3	0.0
P1	-24.0	-14.3	-11.6	-11.5	-6.8	-2.4	0.3	-1.7	-1.5	-2.2	-2.2	13.0	10.8	12.3	12.4	9.3	9.4	3.5	3.8
P1'	-2.0	-0.0	-0.9	-0.4	0.3	-0.8	0.4	0.4	0.1	0.2	0.1	-5.0	-0.1	-0.4	-0.4	1.2	1.6	0.2	-7.9
P2'	-3.5	-1.0	-2.1	-2.8	-0.1	-2.1	-0.2	-0.3	0.2	-0.8	-0.4	0.5	0.5	0.2	-0.1	0.2	1.1	0.5	-1.7
P3'	-0.4	0.0	-0.0	-0.2	0.2	-0.1	0.2	0.0	0.1	0.0	0.2	0.1	-0.3	-0.1	-0.1	0.9	2.1	-5.0	-0.2
P4'	-0.3	0.2	-0.1	-0.2	0.2	0.1	-0.0	0.3	0.1	0.4	0.3	-0.2	-0.2	-0.4	-0.1	0.9	1.5	-0.1	0.3
	T	G	E	S	V	D	H	A	I	N	Q	W	L	F	Y	K	R	M	P

Fig. S3. Unfolded state model parameters and goodness of fit. (A) Parameters for the fitted unfolded state model position-specific scoring matrix (PSSM) for trypsin (eqn. 18). Positive values indicate faster proteolysis, and therefore lower predicted EC_{50} values. Note that the assay does not identify a specific cleavage site (only the global EC_{50} for each sequence is observed), but the training protocol positions the strongest specificity determinants in the center of the 9-residue-length window, which we labeled "P1" for both proteases based on prior knowledge. For visualization purposes, we subtracted the modal value of each row from all entries in that row, producing a mode of zero (kernel density estimation was used to compute the mode of the continuous distribution). Amino acids were clustered using complete linkage hierarchical clustering, with the Euclidian distance between their P5-P4' parameters as the distance metric. Well-known features of trypsin specificity are recovered from our data, including the preference for R/K at P1 (with R favored over K), the inhibitory effect of P at P1' and P3, the favorability of A, M, and V at P1' and P2, and the unfavorability of D and E especially at P2' and P2 (*79*). **(B)** Overall agreement between the model and the data is shown as a 2D histogram counting the number of sequences with a given predicted (x-axis) and actual (y-axis) EC_{50} value, colored on a log scale. Each plot is annotated at the top with the number of scrambled sequences used in training, the overall r^2 between the model predictions and the data, and the root-mean-squared error (RMSE) between the $\log_{10} EC_{50}$ values from the model and the data. **(C, D)** As above, for chymotrypsin. Well-known features of chymotrypsin specificity are again recovered, including the preference for F/Y/W and (less strongly) L at P1, the inhibitory effect of P at P3, P1', and P2', the favorability of K/R at P1' and P3', and the general unfavorability of D and E (*80–82*).

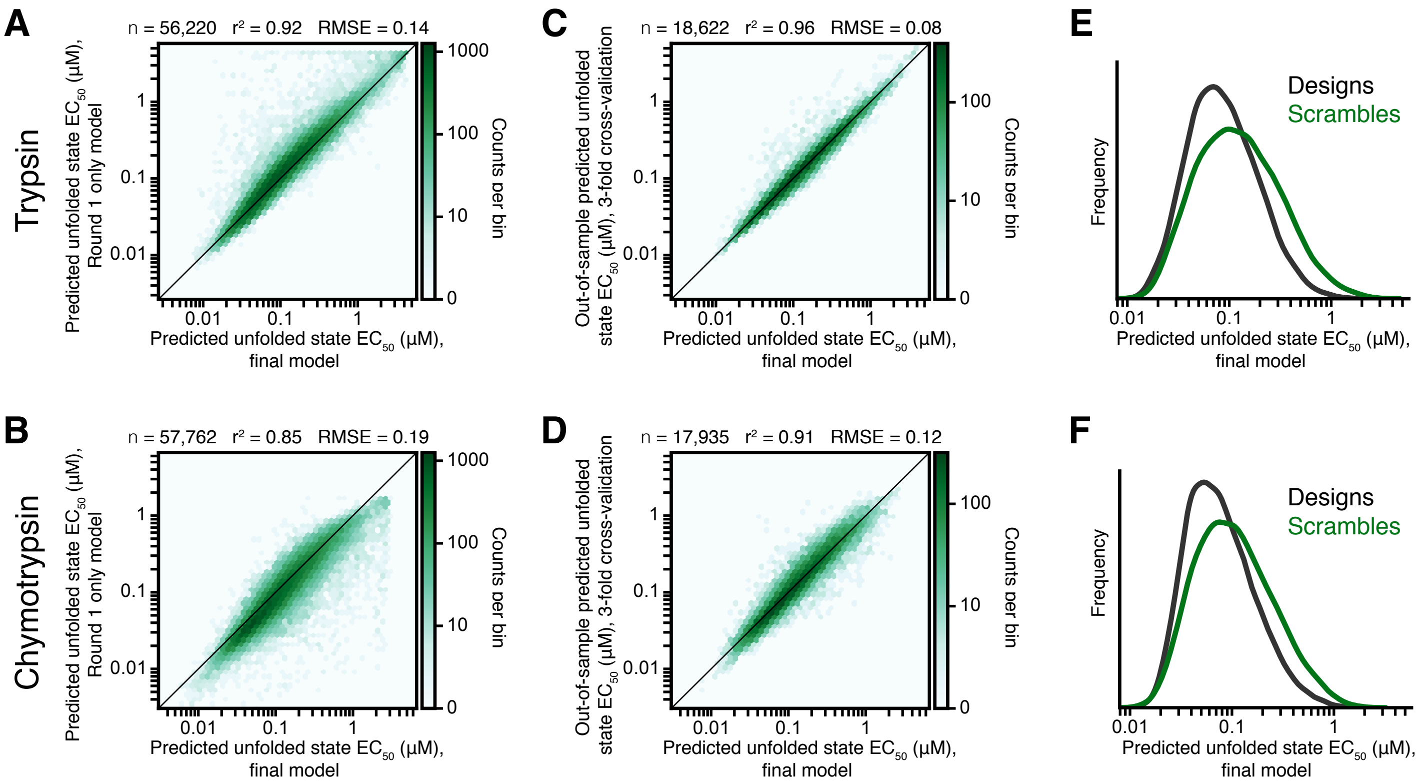


Fig. S4. Training set dependence and bias in the unfolded state model. (**A, B**) Different training protocols for the unfolded state model lead to very similar EC_{50} predictions for all ~55,000 sequences. In the standard protocol (used for all presented data analysis), all ~18,000 scrambled sequences from Rounds 1-4 are used as training data (x-axis); in the alternate protocol, only the ~6,500 scrambled sequences from Round 1 are used as training data (y-axis). Agreement between the training protocols is shown using a 2D histogram counting the number of sequences with a given unfolded state EC_{50} under each protocol, colored on a log scale. (**C, D**) As before, predicted unfolded state EC_{50} values are very similar under two different training protocols. In the standard protocol, all ~18,000 scrambled sequences are used as training data (x-axis). Alternately, three different models for each protease are individually trained; each model excludes a different ⅓ of the data from the training set, and each sequence's unfolded state EC_{50} is predicted using the model for which that sequence is out-of-sample (y-axis). The similarity of the predictions regardless of whether sequences are in-sample or out-of-sample suggests minimal overfitting. (**E, F**) Designed sequences and scrambled sequences have different distributions of predicted unfolded state EC_{50} values. While this could result from sequence effects alone, it more likely indicates some overfitting of the scrambled sequence EC_{50}s (see *Methods: Unfolded state model*). However, the overall magnitude of the difference is small: the median predicted unfolded state trypsin EC_{50} for scrambled sequences is 0.16 stability units (a factor of $10^{0.16}$ in EC_{50}) higher than the median predicted for designed sequences (**E**). This difference is 0.15 stability units for chymotrypsin (**F**).

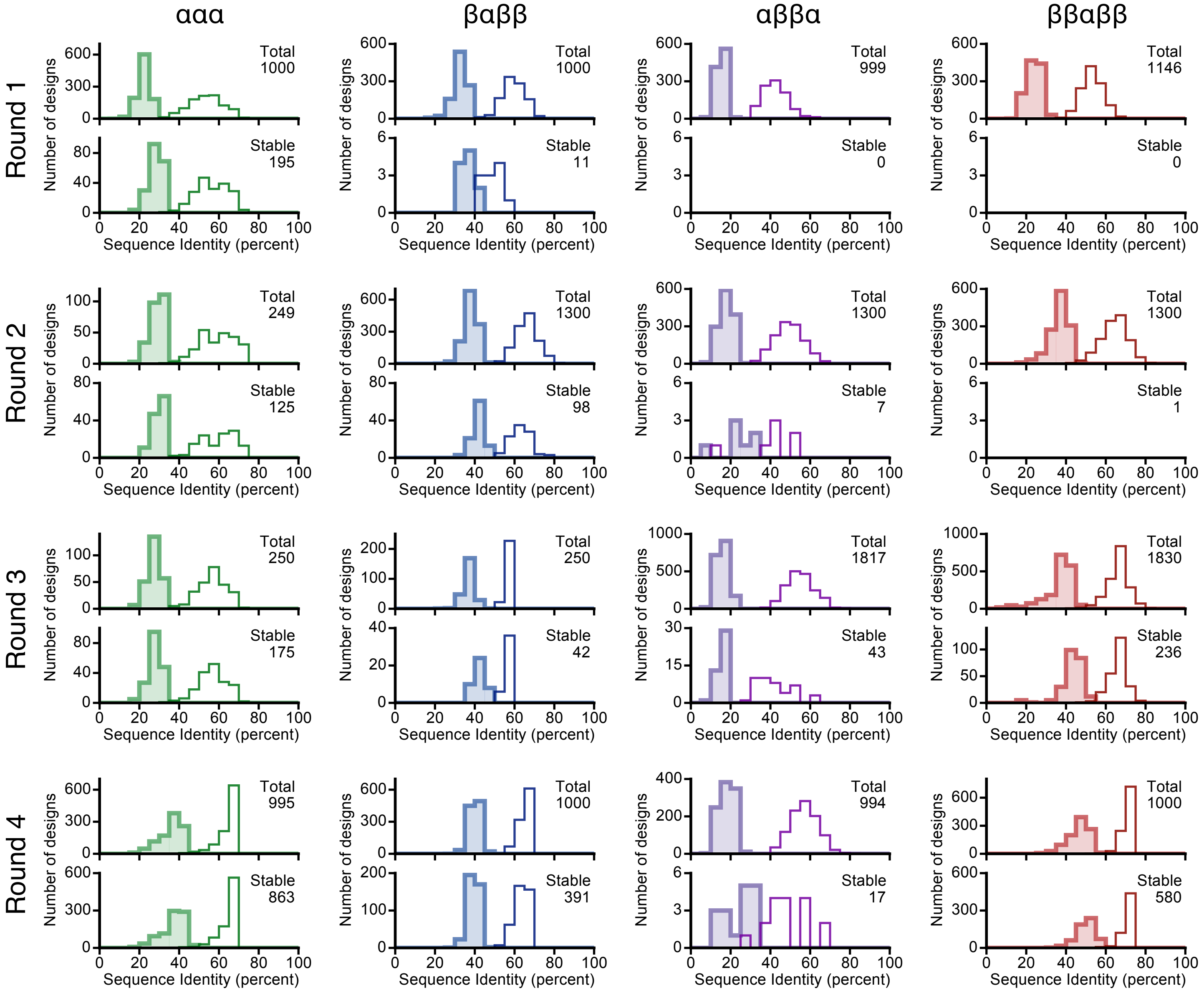
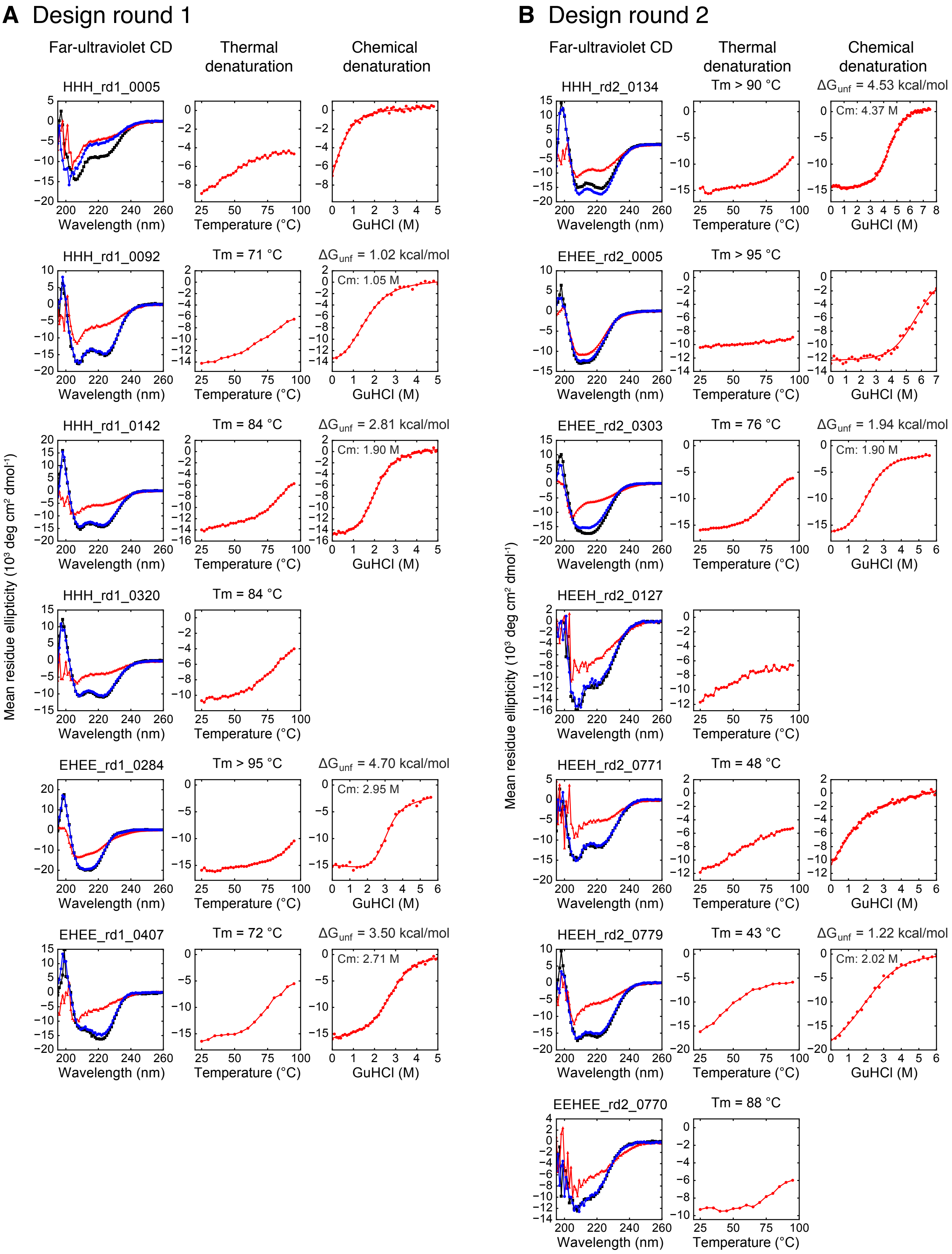


Fig. S5. Sequence identity of designs. We computed the median and maximum sequence identity between each design and all other designs of the same topology and design round. For a given design round (row) and topology (column), the distributions of these identities are shown as the upper shaded (median identities) and outlined (maximum identities) histograms. We also computed equivalent median and maximum identities within the set of successful designs (stability score > 1.0); these are shown as the lower histograms. Bins are five percentage points wide. Sequence identities were computed without any alignment (i.e. from raw sequences); however, all sequences within a topology are identical in length except ββαββ sequences. For ββαββ designs, identity was computed for residues 1 to *n*, where *n* is the length of the shorter design. Upper plots are labeled with the total number of designs tested. These differ slightly from the text in some cases (e.g. 994 vs. 1,000) because a small number of designs included Cys residues that were intended to be prohibited; these were removed from this and all other analysis. Lower plots are labeled with the number of stable designs.



(continued on next page)

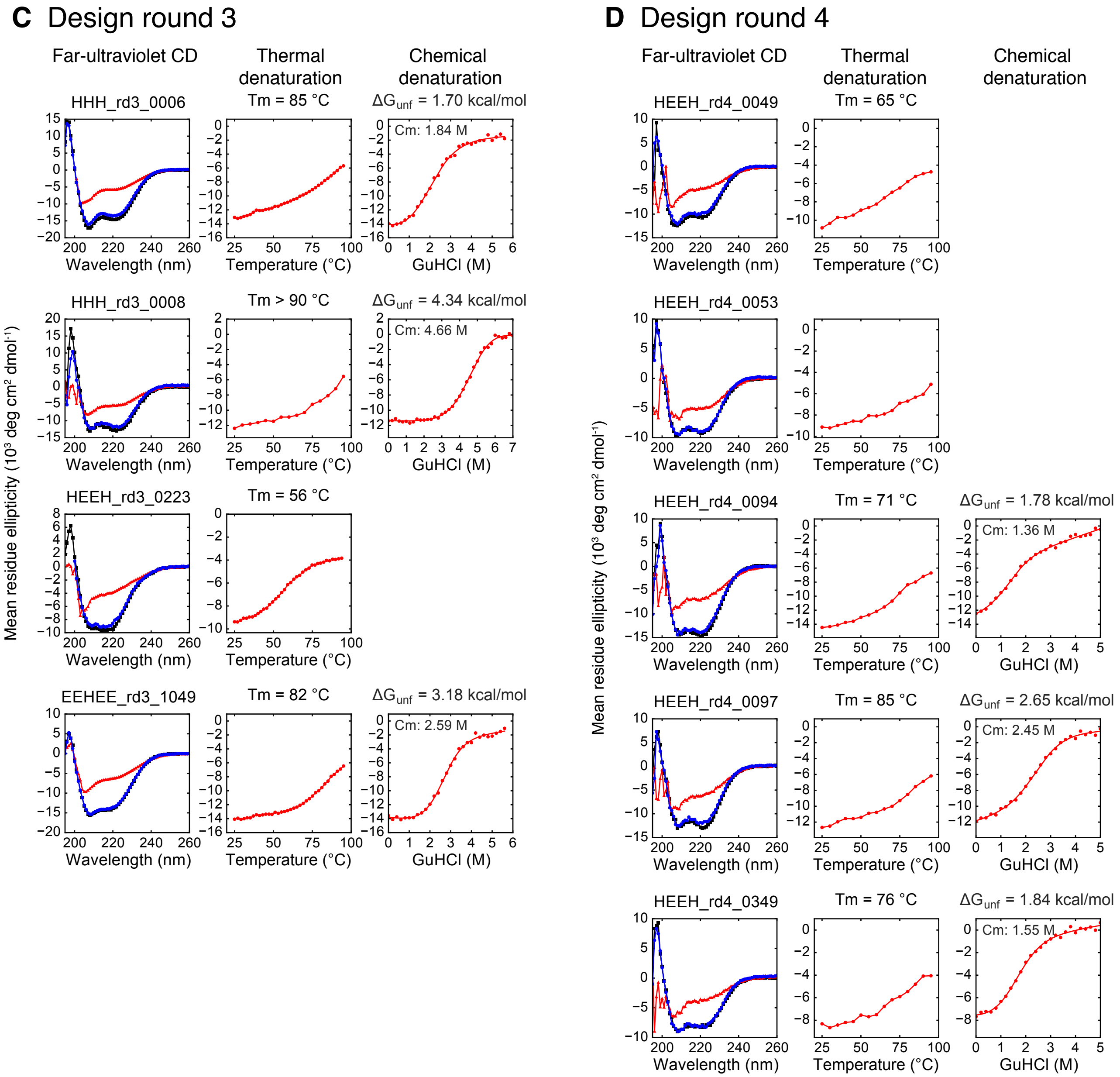


Fig. S6. Characterization of purified designs by circular dichroism (CD) spectroscopy. For each protein, we show (left panel) the far-ultraviolet CD spectrum at 25°C (black), 95°C (red), and 25°C following melting (blue), (center panel) thermal melting curves measured by CD at 220 nm, and (right panel) chemical denaturation curves in GuHCl measured by CD at 220 nm at 25°C. Melting temperatures were determined using the derivative of the curve, and unfolding free energies were determined by fitting to a two-state model (red solid line), as described in *Methods: Circular dichroism*. Designs were assayed as either the designed sequence alone or including an additional 21-residue-length unstructured tag at the N-terminus; designs expressed with the pET-28b+ method (see Table S1) included this tag. Chemical denaturation assays were not performed in cases where purified protein yield was low.

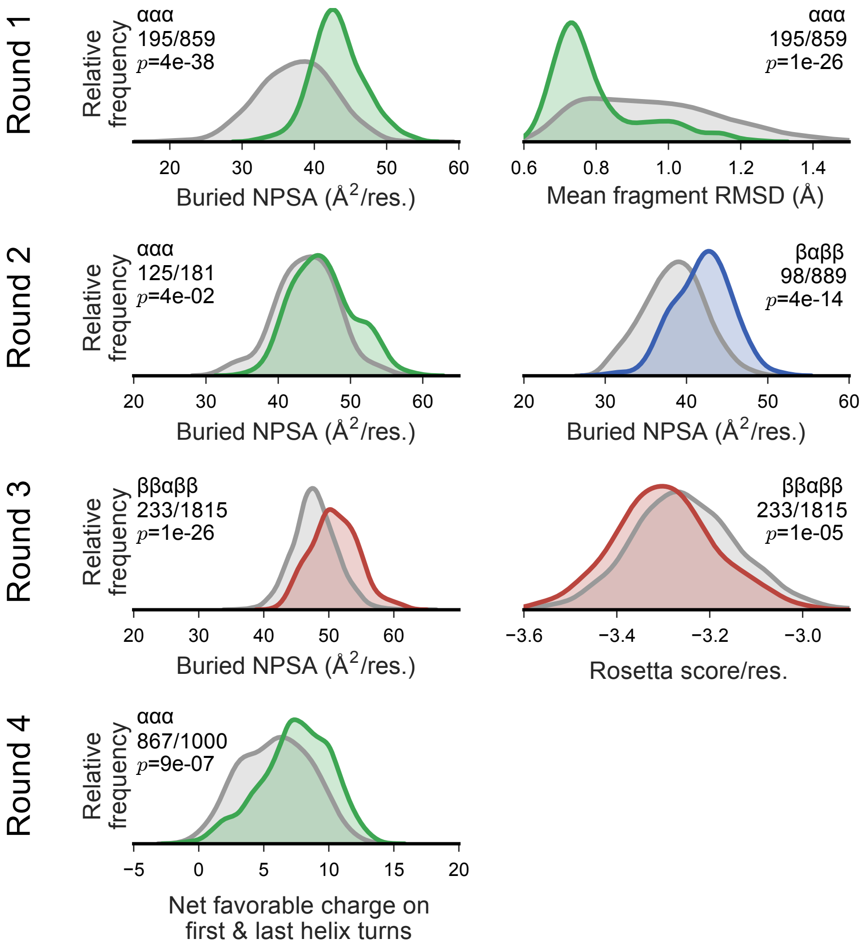


Fig. S7. Feature distributions of stable and unstable designs. To complement the analysis in Fig. 2 B-G, we show the distributions of the structural metrics from Fig. 2 for stable designs (colored distributions, colors correspond to the different topologies) and unstable designs (grey distributions). Distributions are plotted as kernel density estimates; definitions of the metrics analyzed are given in Fig. 2. As in Fig. 2, each panel displays data for designs of a specific topology from a specific design round, and panels are annotated with their topology, the number of stable designs and total number of designs tested, and the *p*-value for the null hypothesis that the scores of the stable and unstable designs were drawn from the same underlying distribution, computed using the Kolmogorov-Smirnov 2-sample test. These *p*-values assume that all designs are independent; see Fig. S5 for the sequence identities between designs.

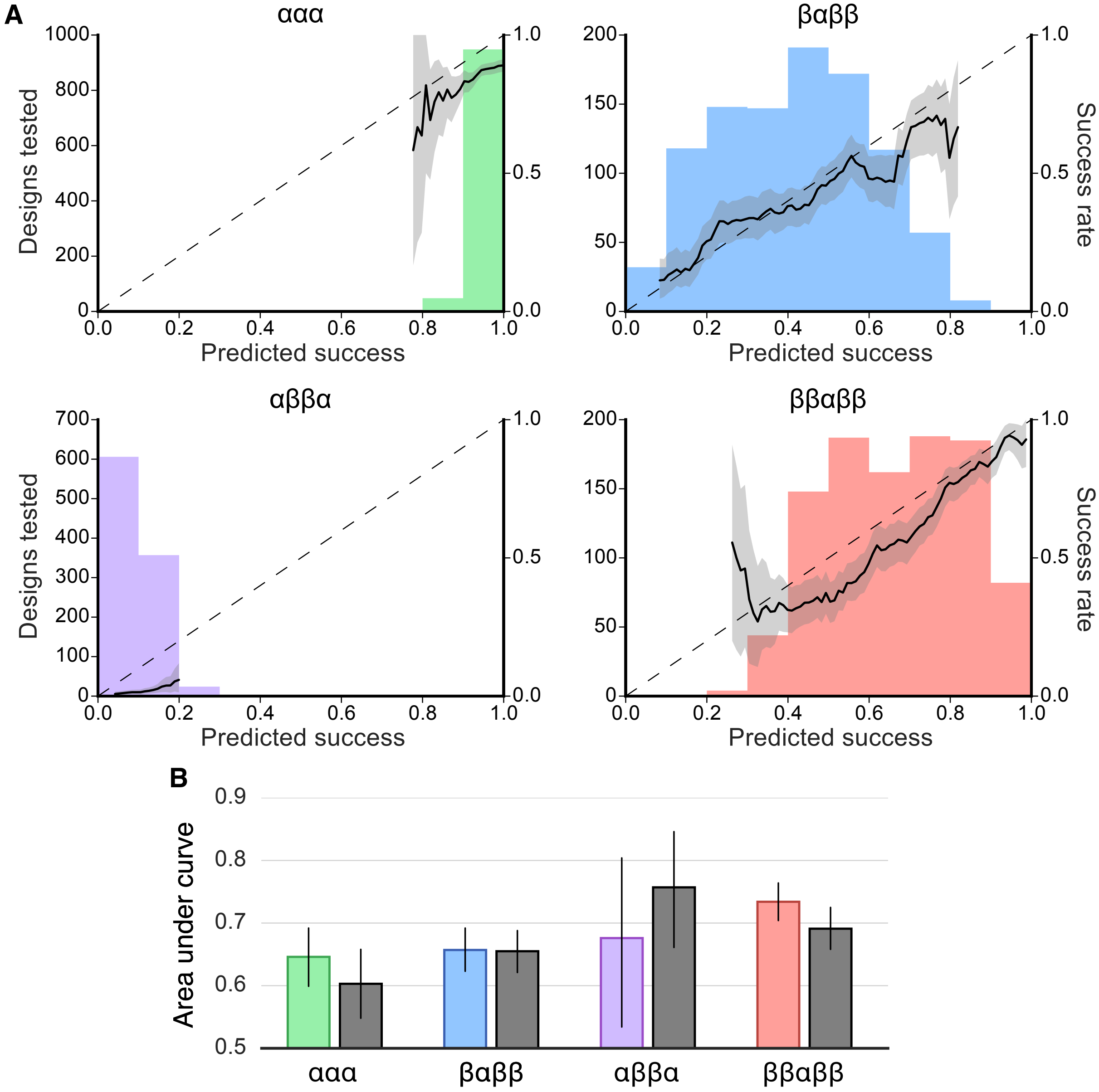


Fig. S8. Accuracy of the logistic regression models for each topology. (**A**) To complement the analysis in Fig. 2I, we again plot the success rate of designs from Round 4 (solid black line, right y-axis) as a function of the predictions (made in advance) of the logistic regression models used to select designs for testing (x-axis). Unlike in Fig. 2I, success rates are computed for each topology individually instead of averaged over all four topologies. Success rates were computed as moving averages over bins of width 0.1; the 95% confidence interval of the success rate from 1,000 bootstrapping trials is indicated by grey shading. The dashed line represents perfect agreement between predicted success rates and actual success rates. Each panel also displays a histogram counting the number of designs of that topology at each level of predicted success rate (left y-axis). (**B**) A single logistic regression trained on round 1-3 data from all topologies provides comparable performance to the topology-specific logistic regressions at ranking Round 4 designs. Performance is quantified using the receiver-operator characteristic area under curve (ROC AUC) metric. For each topology, we show the AUC value of the topology-specific regression as the left colored bar (with its 95% confidence interval from bootstrapping), and show the AUC value and confidence interval of the single overall regression as the right grey bar.

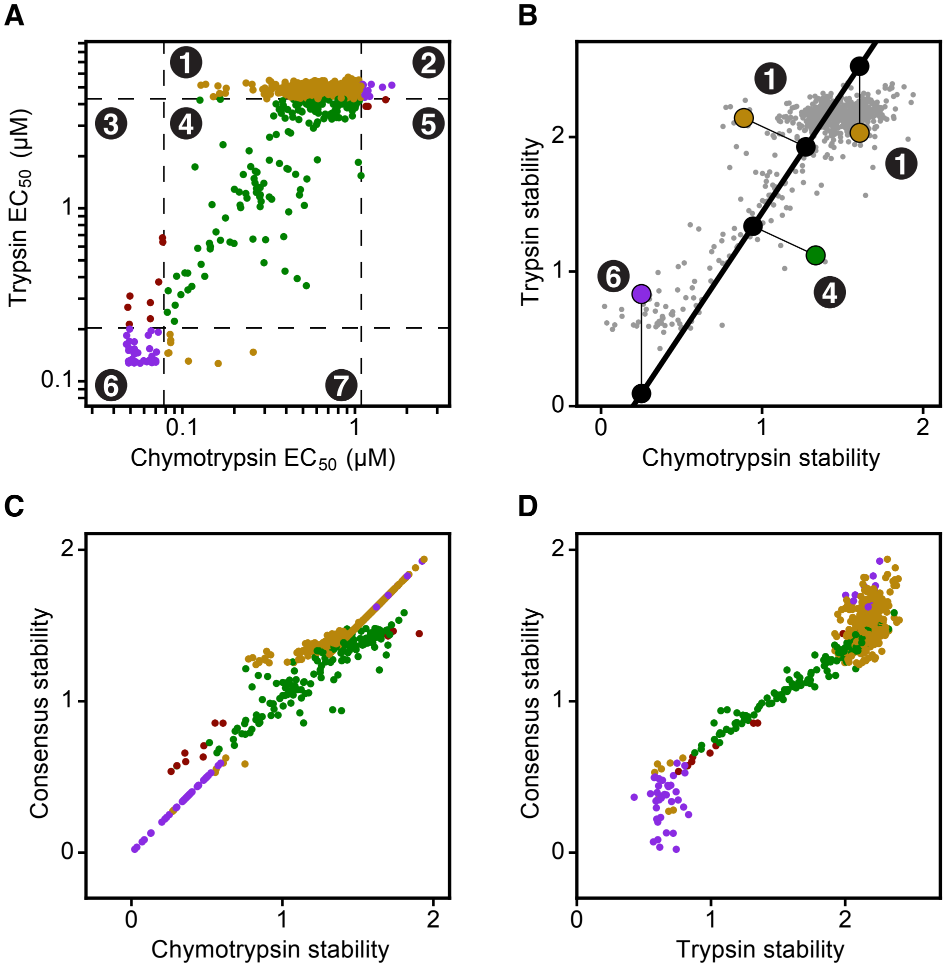


Fig. S9. Determination of consensus stability scores of point mutants from trypsin and chymotrypsin stabilities. We combined the trypsin and chymotrypsin stability measurements for each mutant into an overall consensus stability score using a four-step procedure. This procedure is described in *Methods: Mutational stability effects* and accounts for the different slopes observed between the trypsin and chymotrypsin datasets for each set of mutants, and the differences in reliable dynamic range between the two datasets. Here, we illustrate the procedure for the mutants of the protein EHEE_rd1_0882 (Fig. S10 D). (**A**) First, all mutants are classified by whether their estimated EC_{50} values are below, within, or above the reliable dynamic range of the assay. Mutants are colored according to this classification and the thresholds are indicated by dashed lines; regions of the plot are numbered 1-7 according to whether each region is within the assay's reliable dynamic range. (**B**) Second, the high-confidence measurements from region 4 are used to determine the best-fit line (solid black) between the trypsin and chymotrypsin data by orthogonal distance regression. Each parent protein has a unique best-fit line determined by all the mutants from that protein (in this case, EHEE_rd1_0882). Third, each mutant of a given protein is mapped onto the best-fit line for that protein at one of three positions: the nearest point on the fit line, the point on the fit line at an identical x-coordinate (chymotrypsin stability score), or the point on the fit line at an identical y-coordinate (trypsin stability score). The choice of position for a given mutant is determined by whether the chymotrypsin and trypsin EC_{50} values for that mutant were within the dynamic range of concentrations measured with each protease. EC_{50} values above and

below the dynamic range of concentrations tested have large uncertainties, and our EC_{50} fitting procedure (described in *Methods: EC_{50} estimation from sequencing counts*) estimated these EC_{50}s to be near the limits of our tested dynamic range even when they could be significantly outside the range. We chose the mapped location for each mutant (out of the three possibilities) to avoid artificially high or low stability scores caused by these uncertain EC_{50} values. For example, the highlighted mutant from region 6 (colored in purple, lower left) has EC_{50} values estimated to be below the useful dynamic range with both trypsin and chymotrypsin. Both the chymotrypsin and trypsin EC_{50} values were likely estimated to be too high, so we mapped this mutant onto the best-fit line at the position (among the three options) with the lowest overall stability score, which for this mutant is the position on the best-fit line sharing the x-coordinate (chymotrypsin stability score), shown by a black circle. However, the mutant highlighted in green in Fig. S9B from region 4 was in the confident EC_{50} dynamic range for both proteases, and was therefore mapped onto the best-fit line at the closest position (black circle). Two mutants from region 1 (highlighted in gold) were mapped onto the best-fit line differently from each other. The mutant above and to the left of the fit-line was mapped to the closest position on the best-fit line, because even though that mutant's trypsin EC_{50} lies above the reliable dynamic range and may be spuriously low, inclusion of this high EC_{50} raises the overall consensus stability score for this mutant rather than lowers it. The rightmost mutant highlighted in gold is mapped to the position on the best-fit line sharing the same x-coordinate (directly above the gold point), because inclusion of the trypsin stability score (again, potentially spuriously low) in the mapping might spuriously lower the consensus stability score due to the dynamic range limits of the trypsin assay. Finally, after mapping all points onto the fit line, we used the x-coordinate of each mapped point (the location of that point on the chymotrypsin axis) as the overall consensus stability score for each mutant. (**C**) Consensus stability scores for all mutants, colored by region in panel A and plotted against their chymotrypsin stability scores. The consensus stability scores for the highest- and lowest-stability mutants are determined entirely from the chymotrypsin data because this data covers a wider dynamic range than the trypsin data for this protein, while the stability scores in the middle-range result from the consensus of the chymotrypsin and trypsin data. (**D**) Consensus stability scores for all mutants, colored by region and plotted against their trypsin stability scores. Although the trypsin data lacks the dynamic range to resolve the highest- and lowest- stability mutants, these mutants are still resolved in consensus stability score due to the additional dynamic range provided by the chymotrypsin data.

Fig. S10 A: HHH_rd1_0142

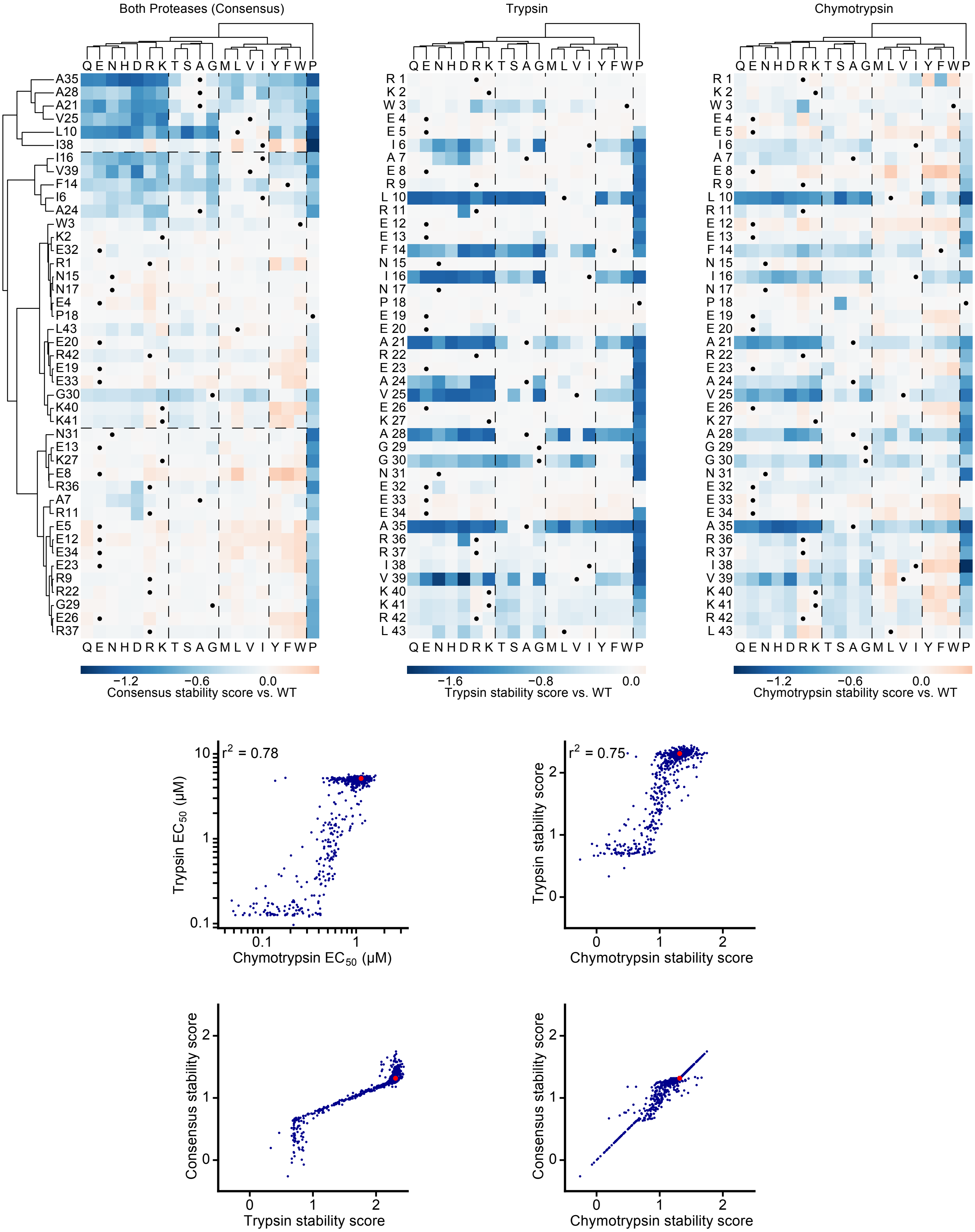


Fig. S10 B: HHH_rd2_0134

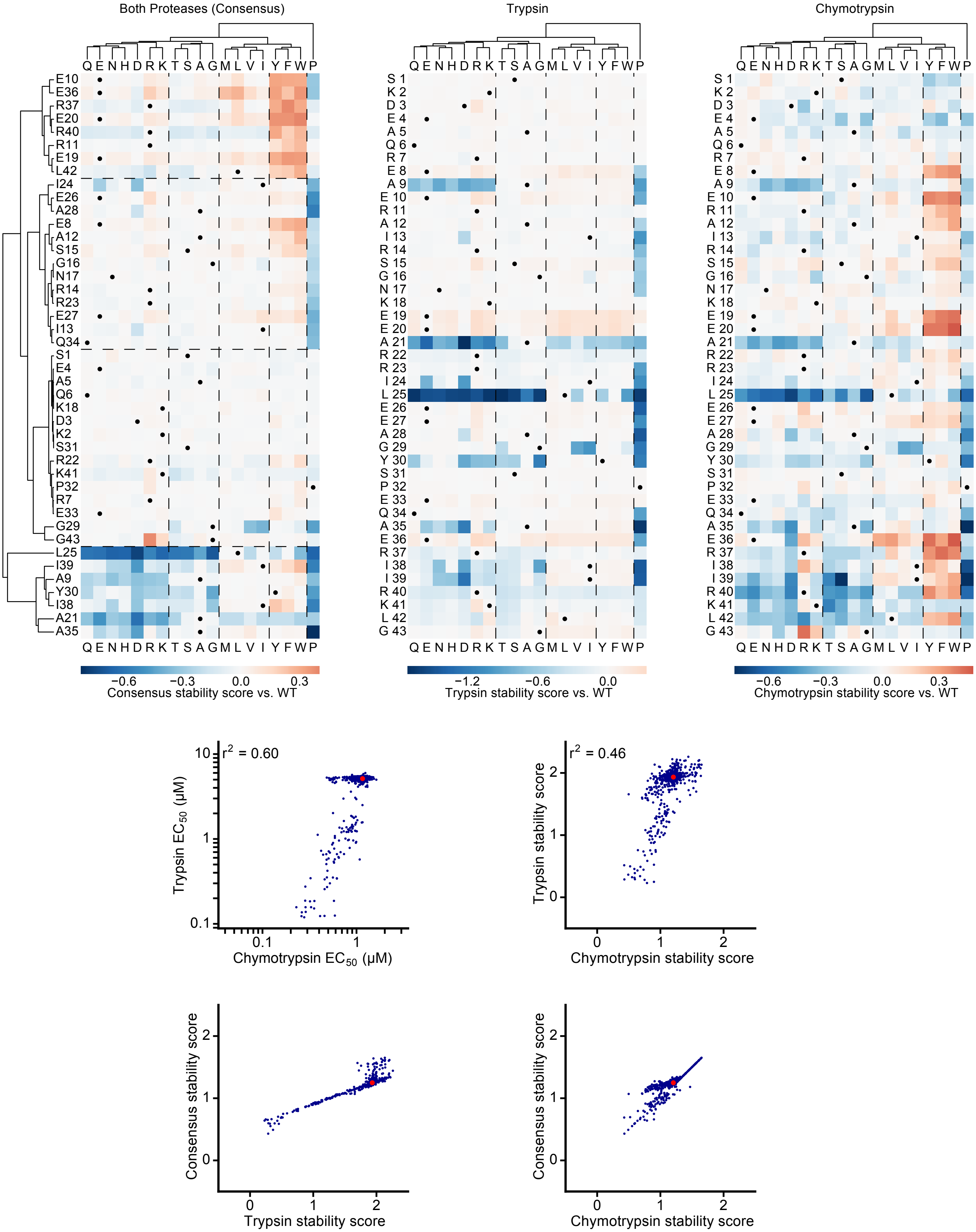


Fig. S10 C: HHH_rd3_0138

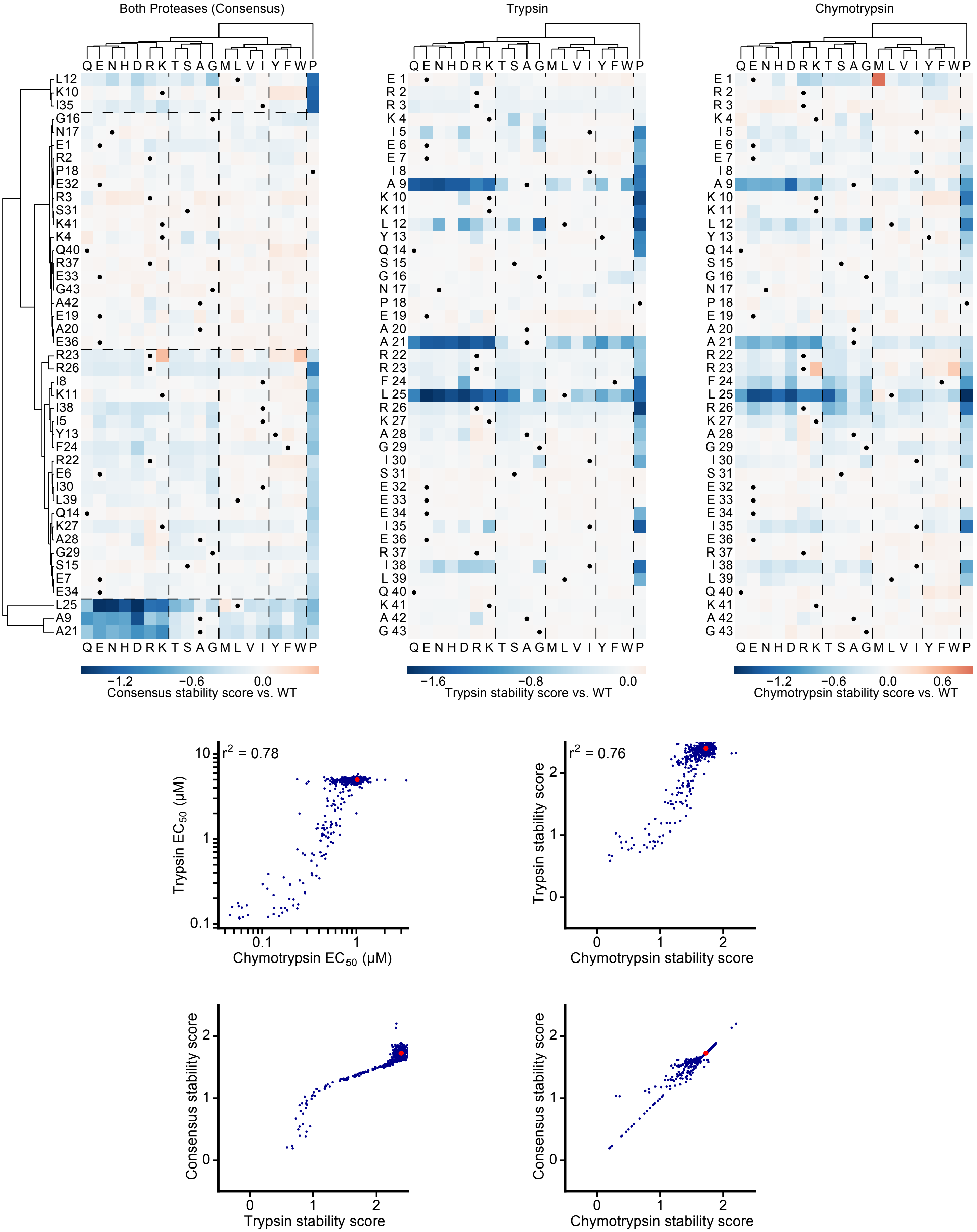


Fig. S10 D: EHEE_rd1_0882

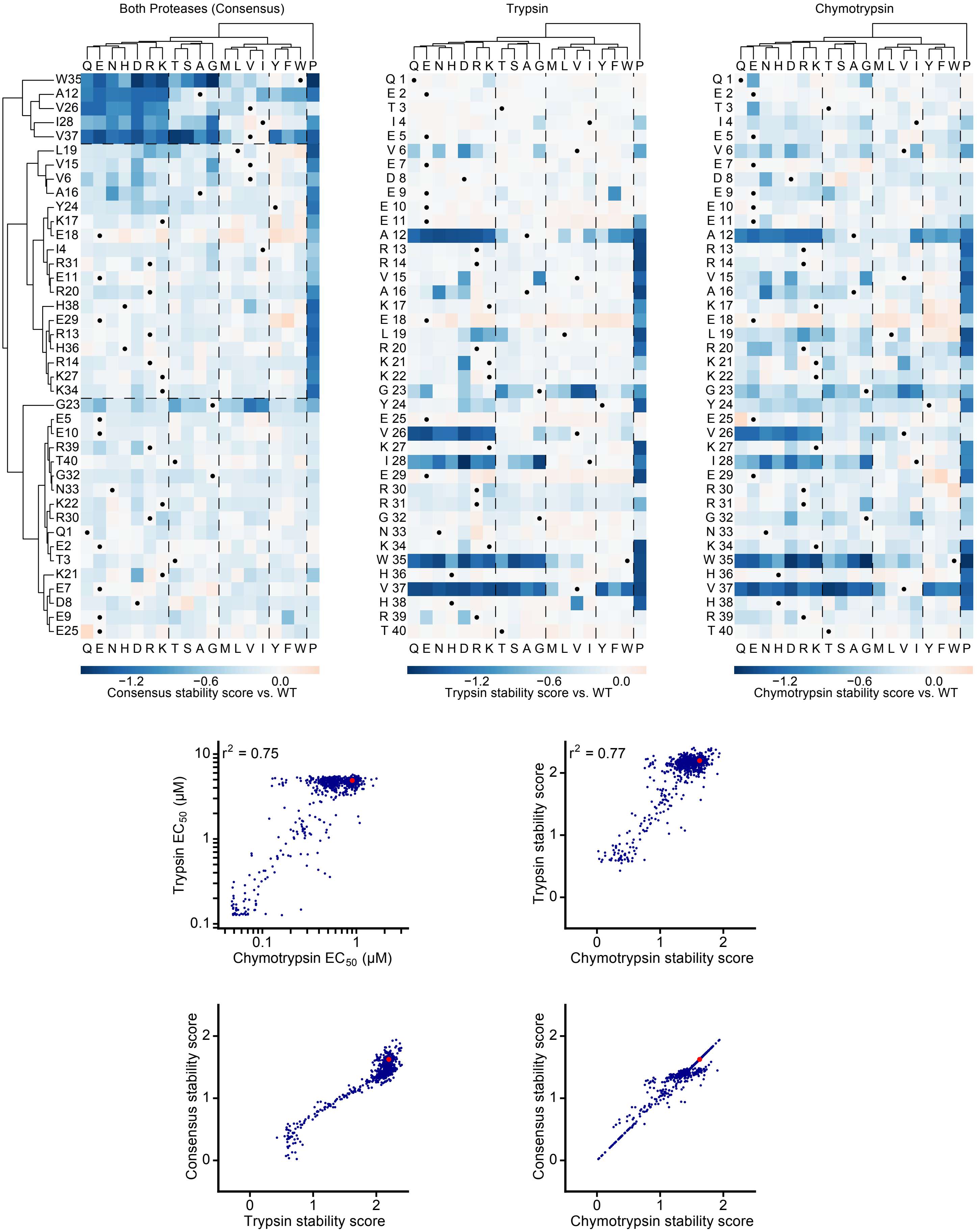


Fig. S10 E: EHEE_rd2_0005

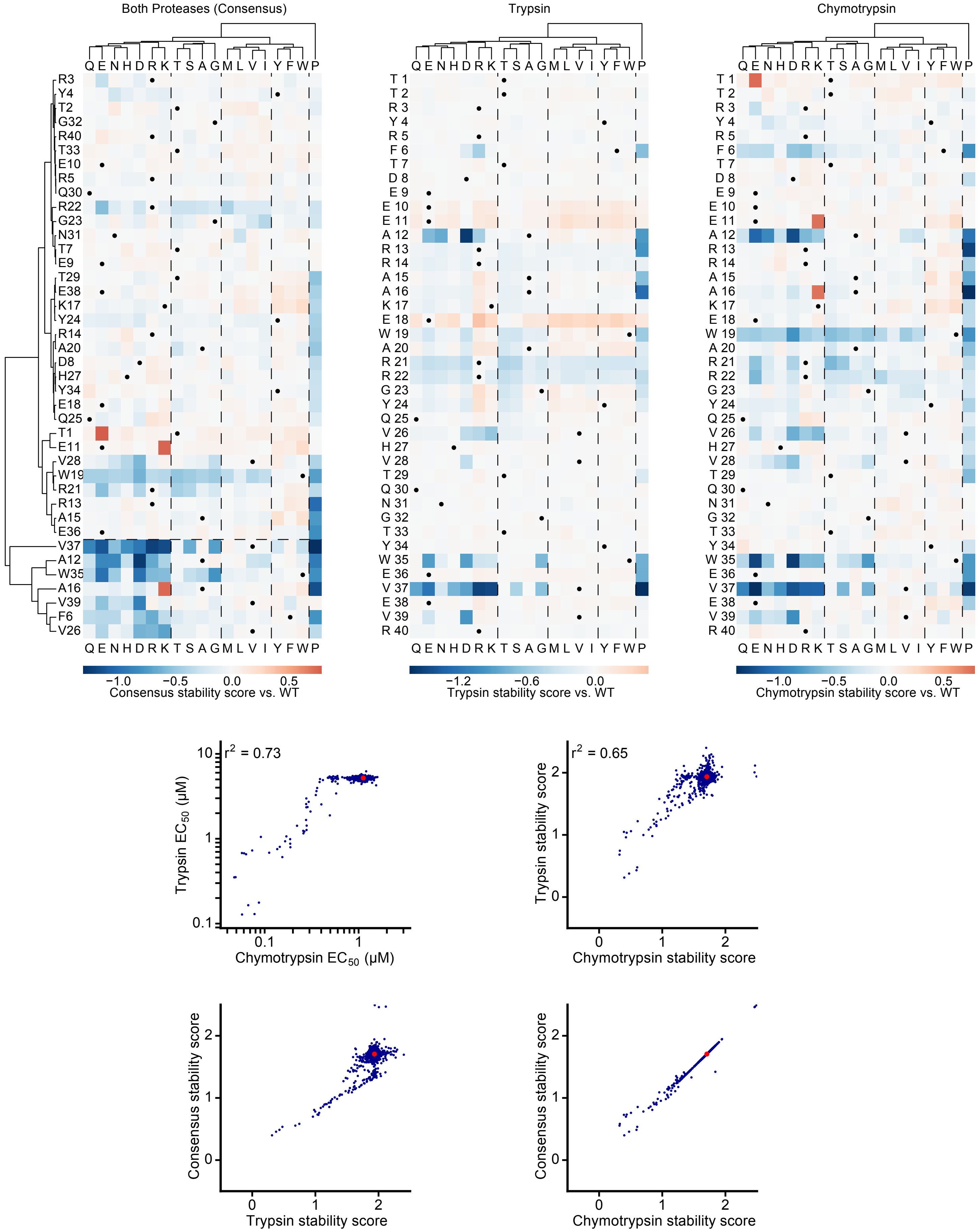


Fig. S10 F: EHEE_rd3_0015

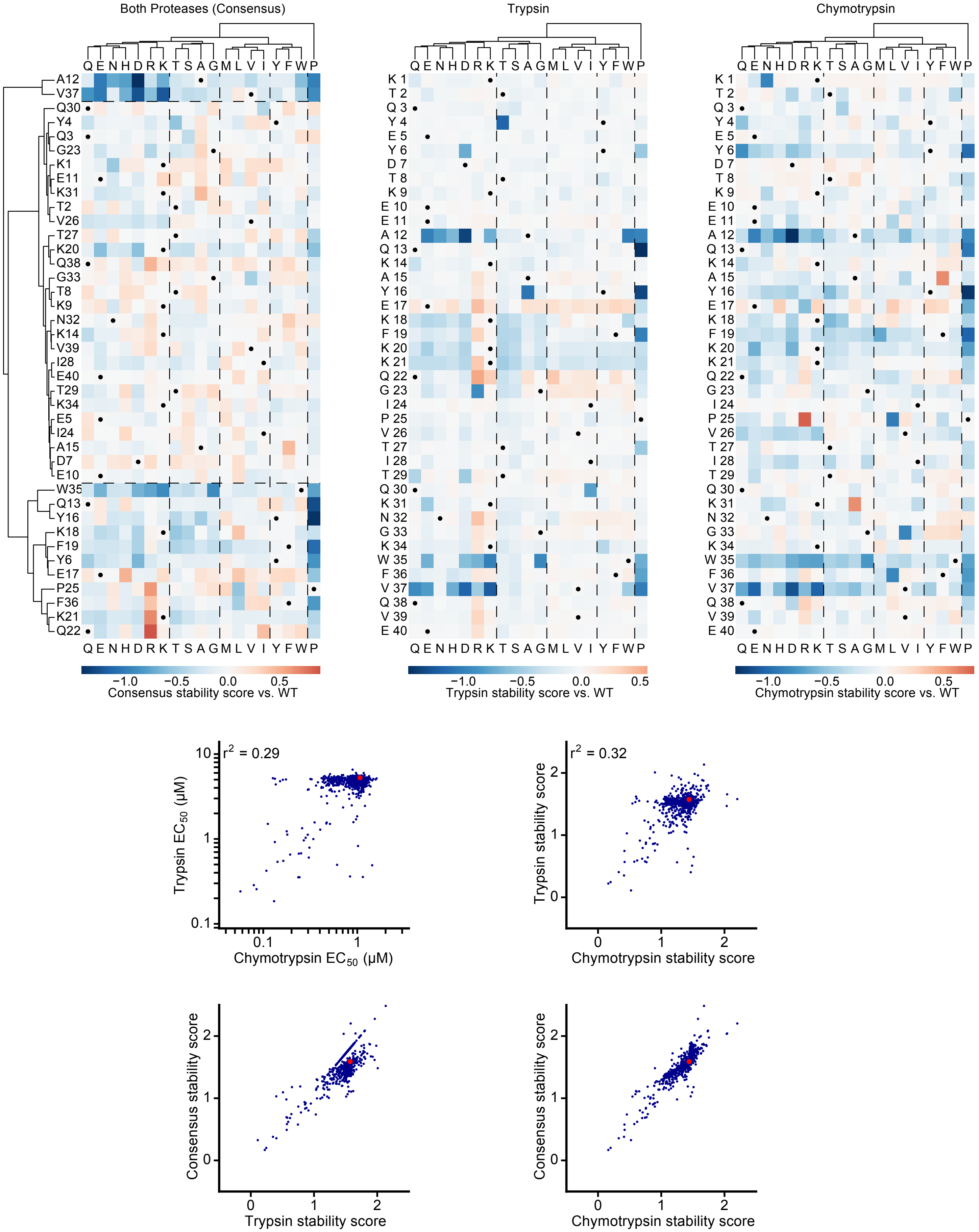


Fig. S10 G: HEEH_rd2_0779

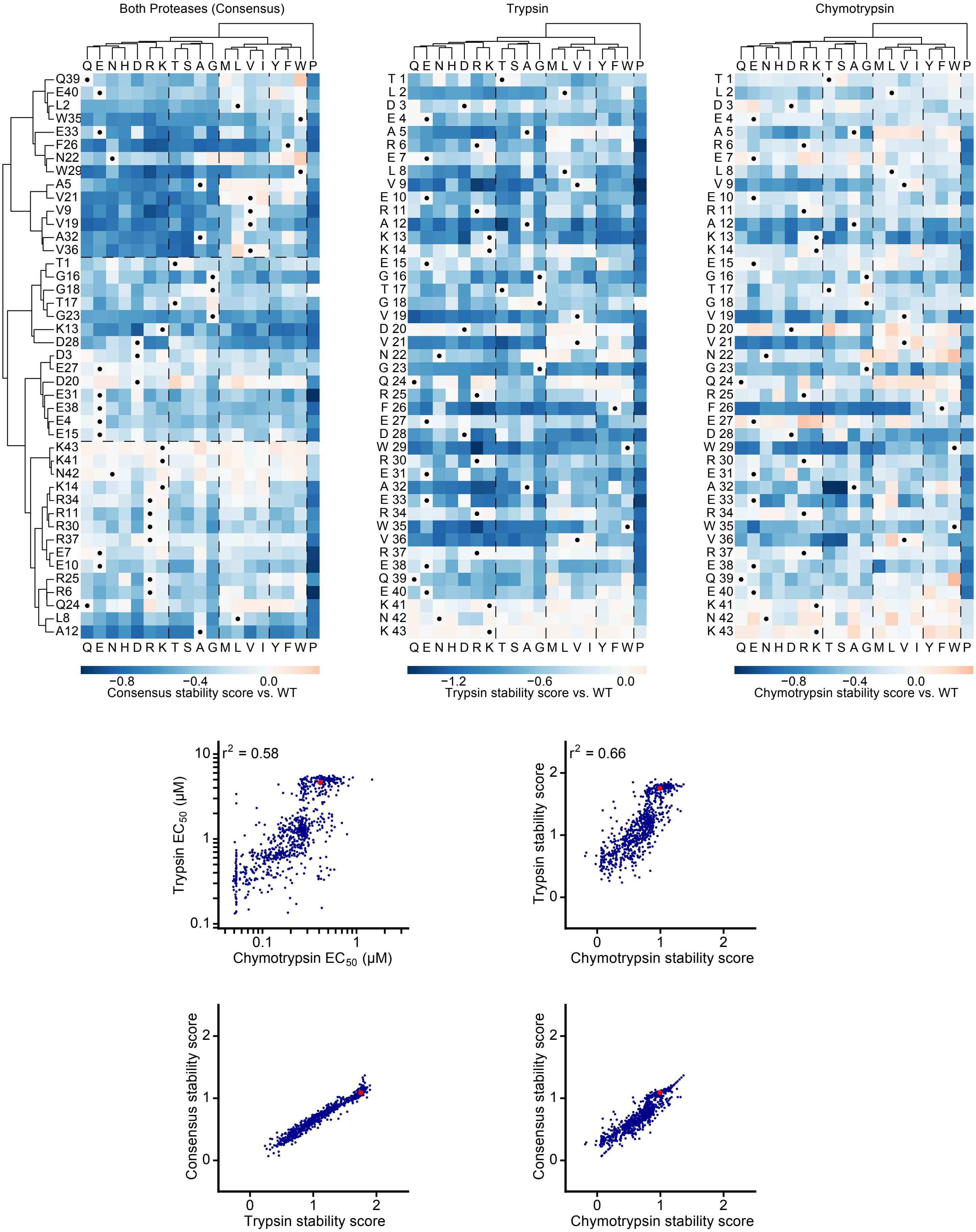


Fig. S10 H: HEEH_rd3_0223

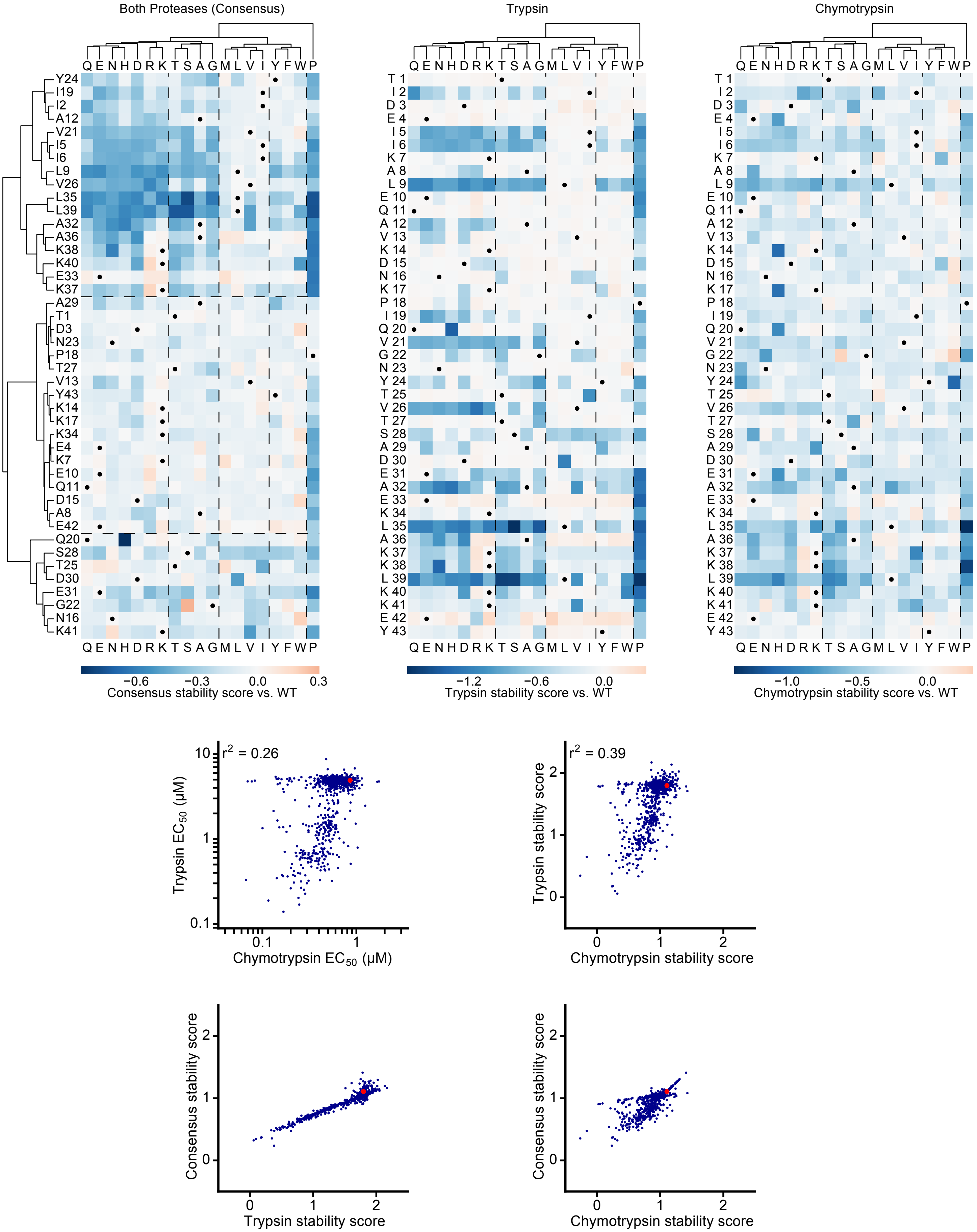


Fig. S10 I: HEEH_rd3_0726

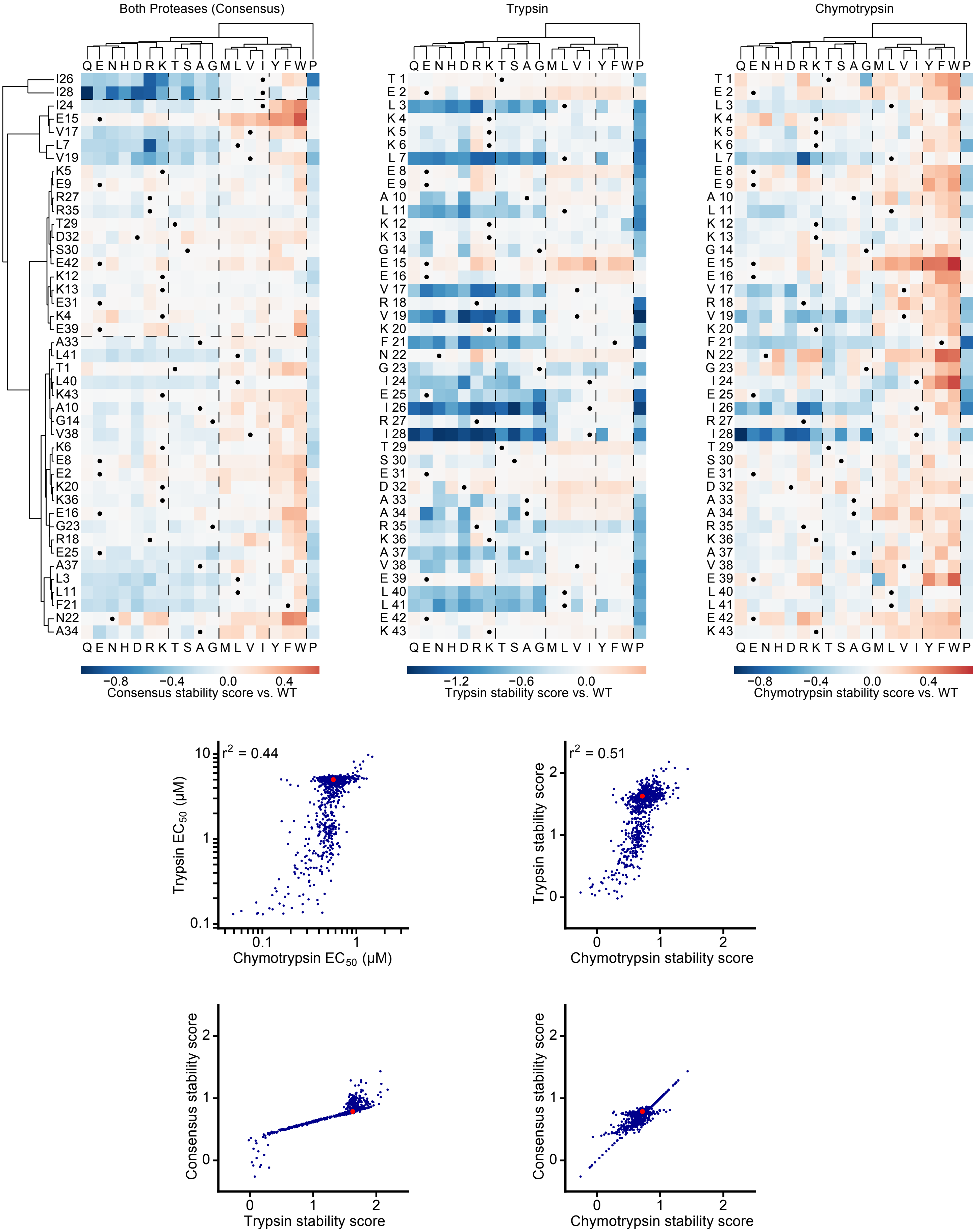


Fig. S10 J: HEEH_rd3_0872

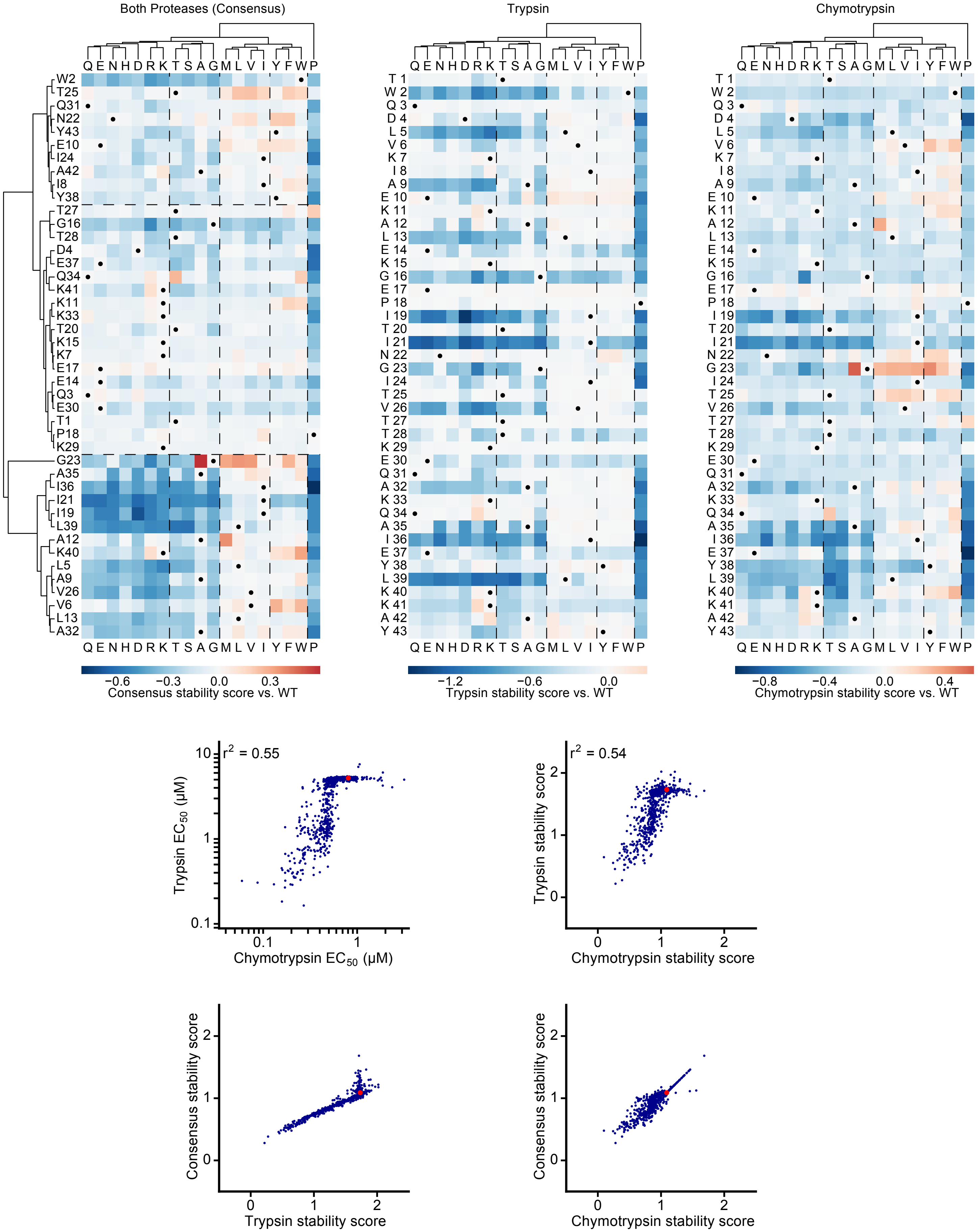
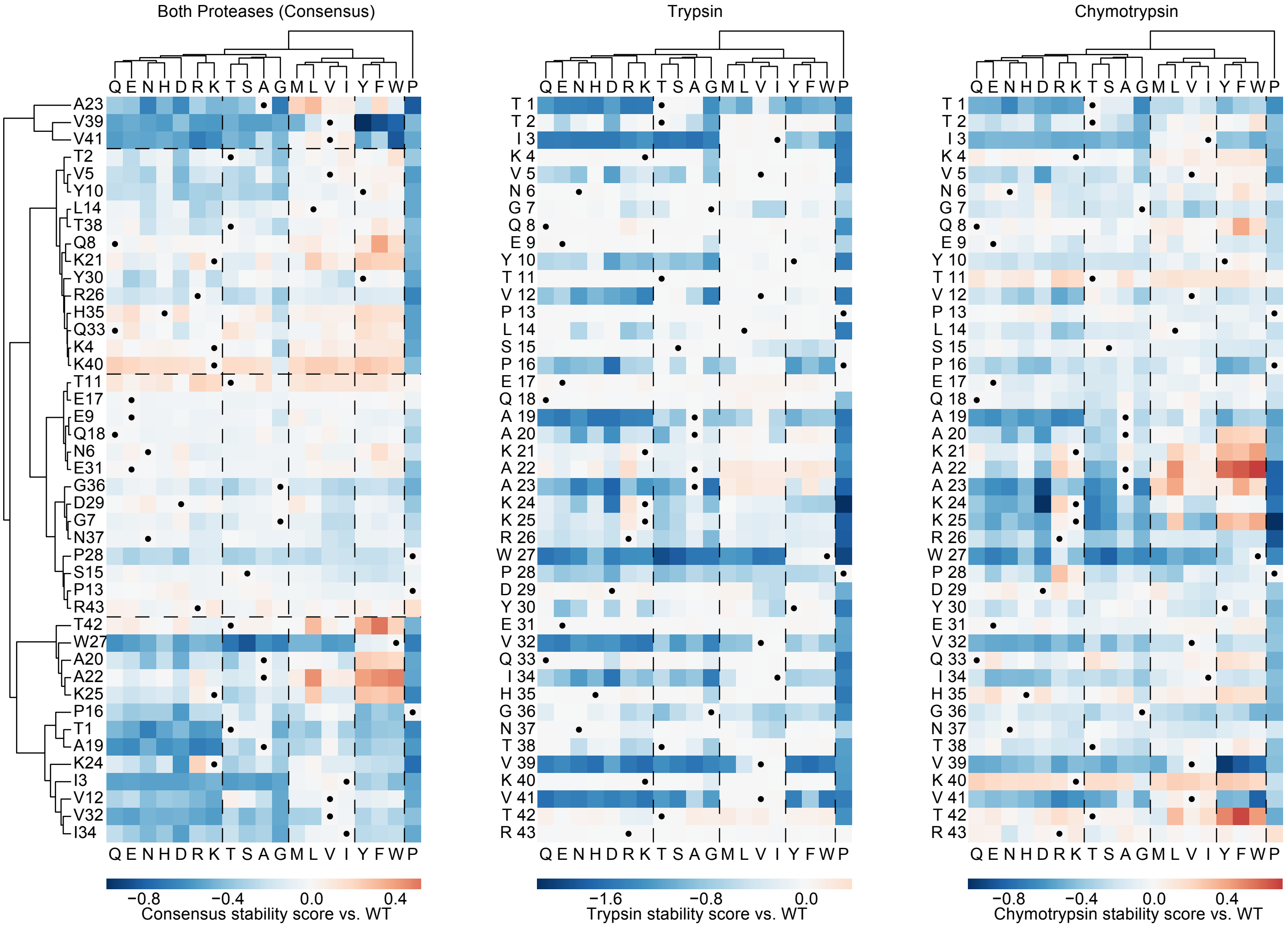


Fig. S10 K: EEHEE_rd3_0037



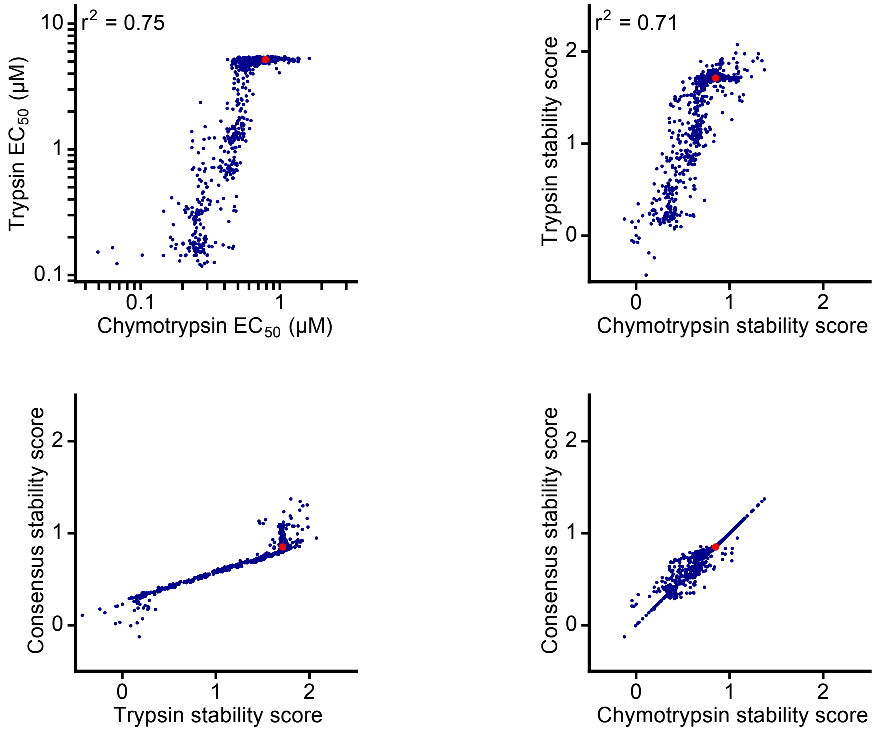


Fig. S10 L: EEHEE_rd3_1498

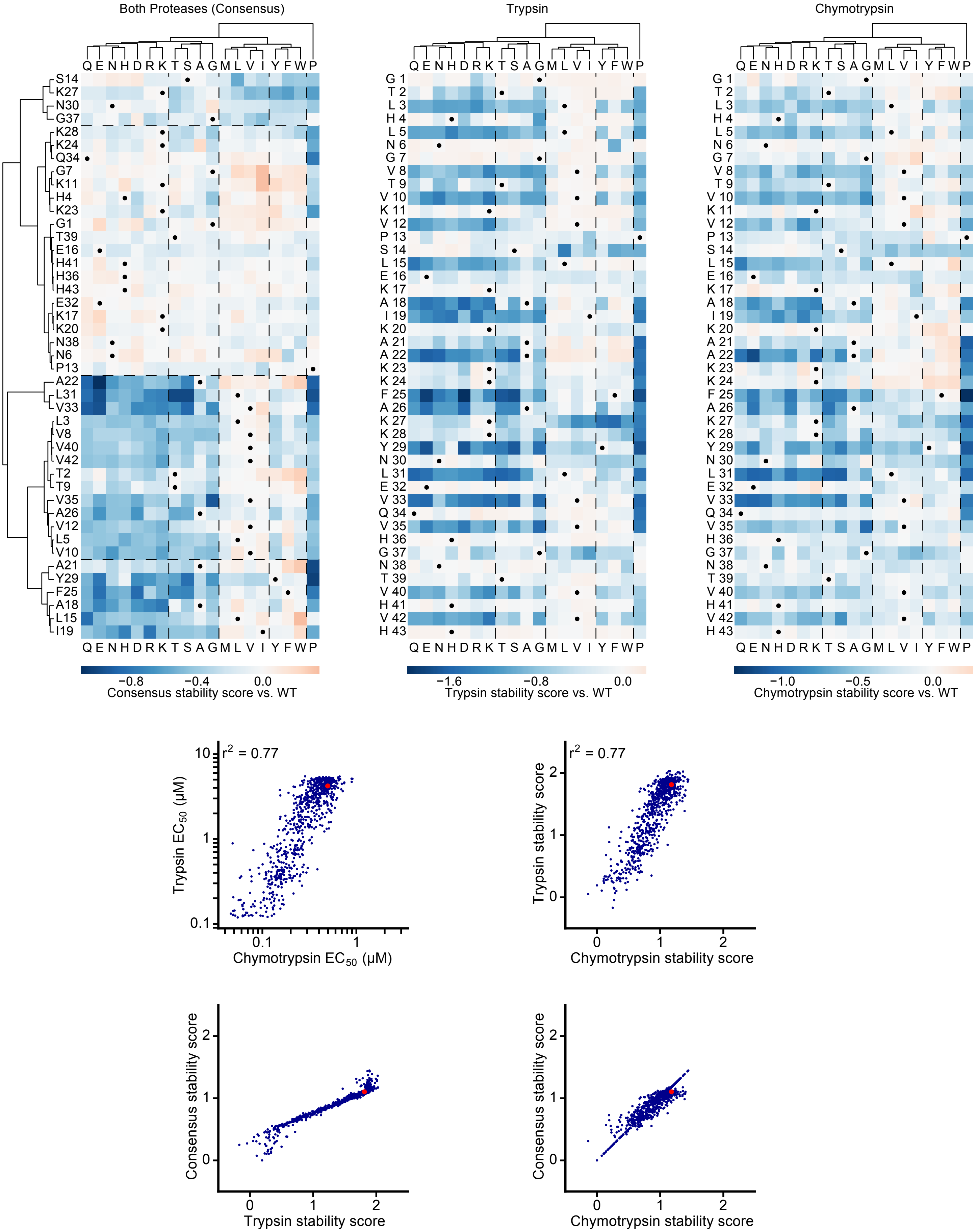


Fig. S10 M: EEHEE_rd3_1702

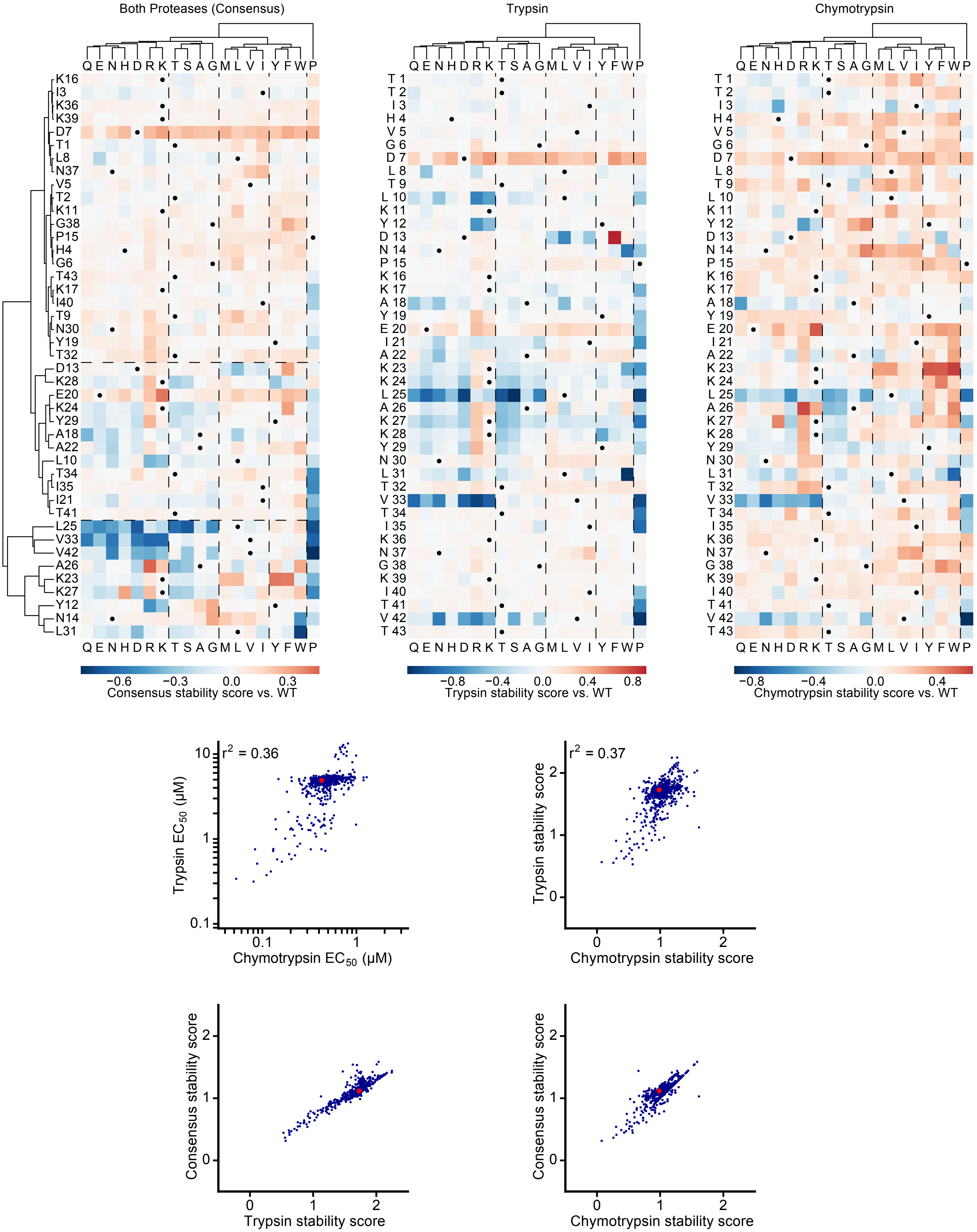


Fig. S10 N: EEHEE_rd3_1716

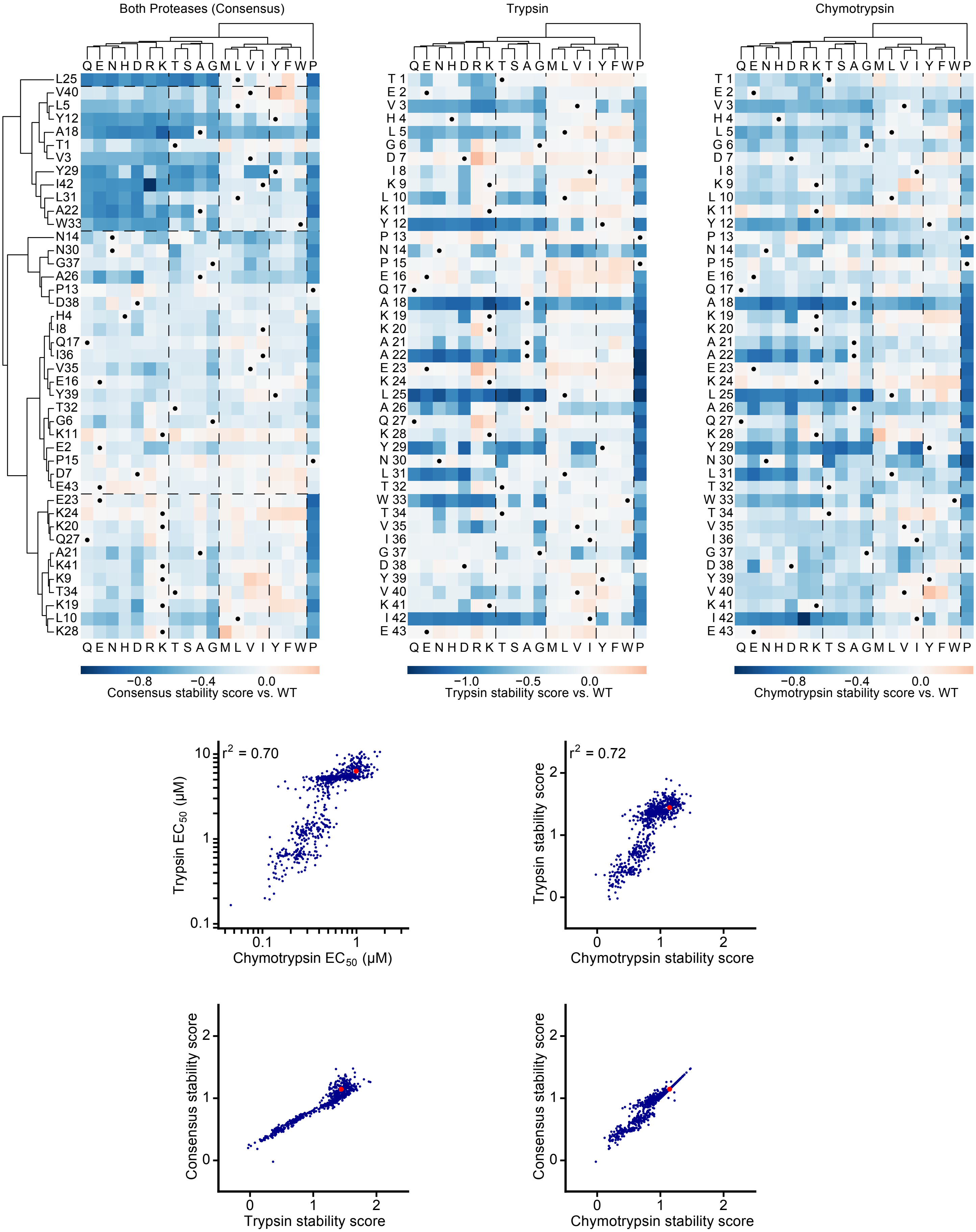


Fig. S10 O: hYAP65 L30K

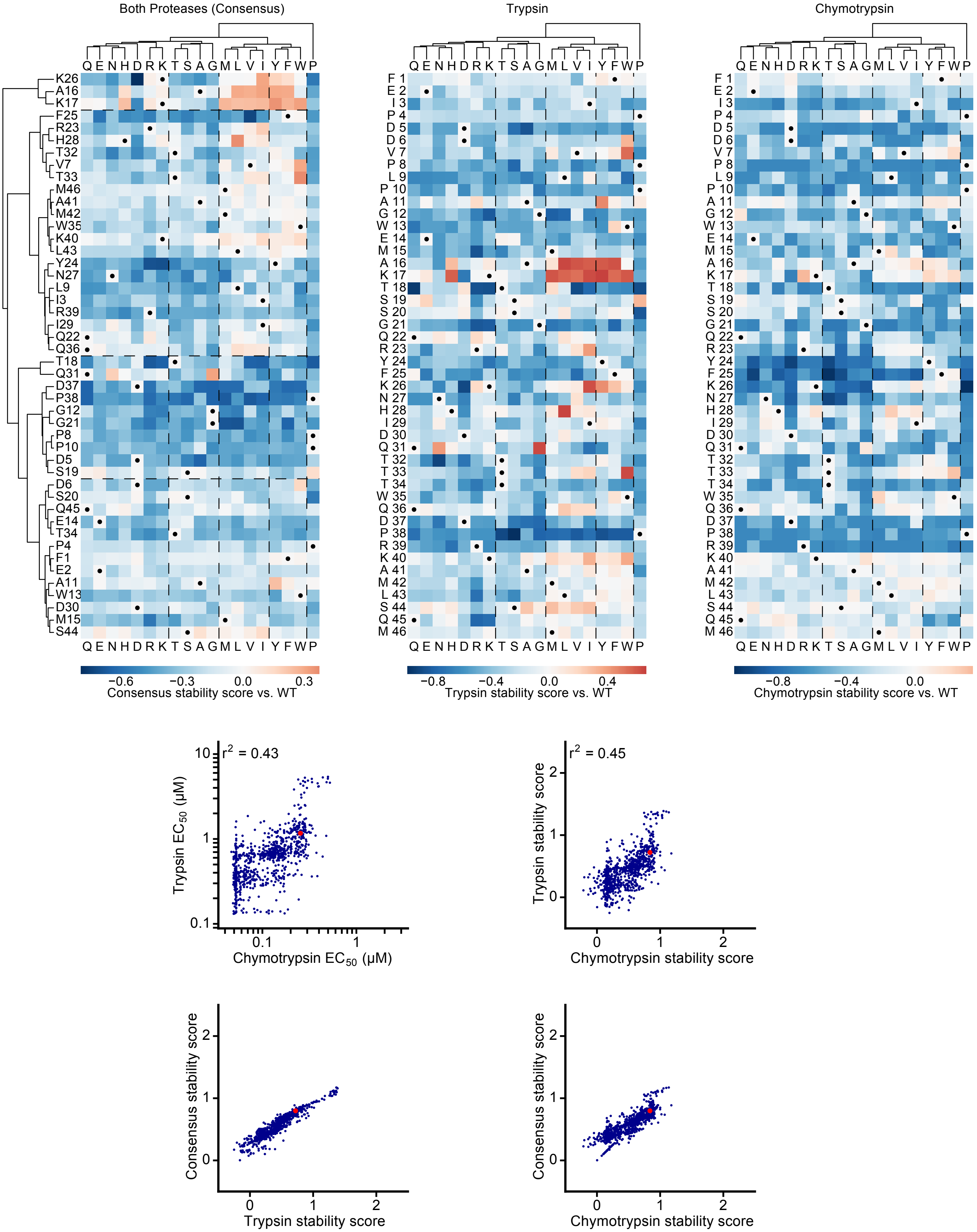


Fig. S10 P: pin1 WW-domain

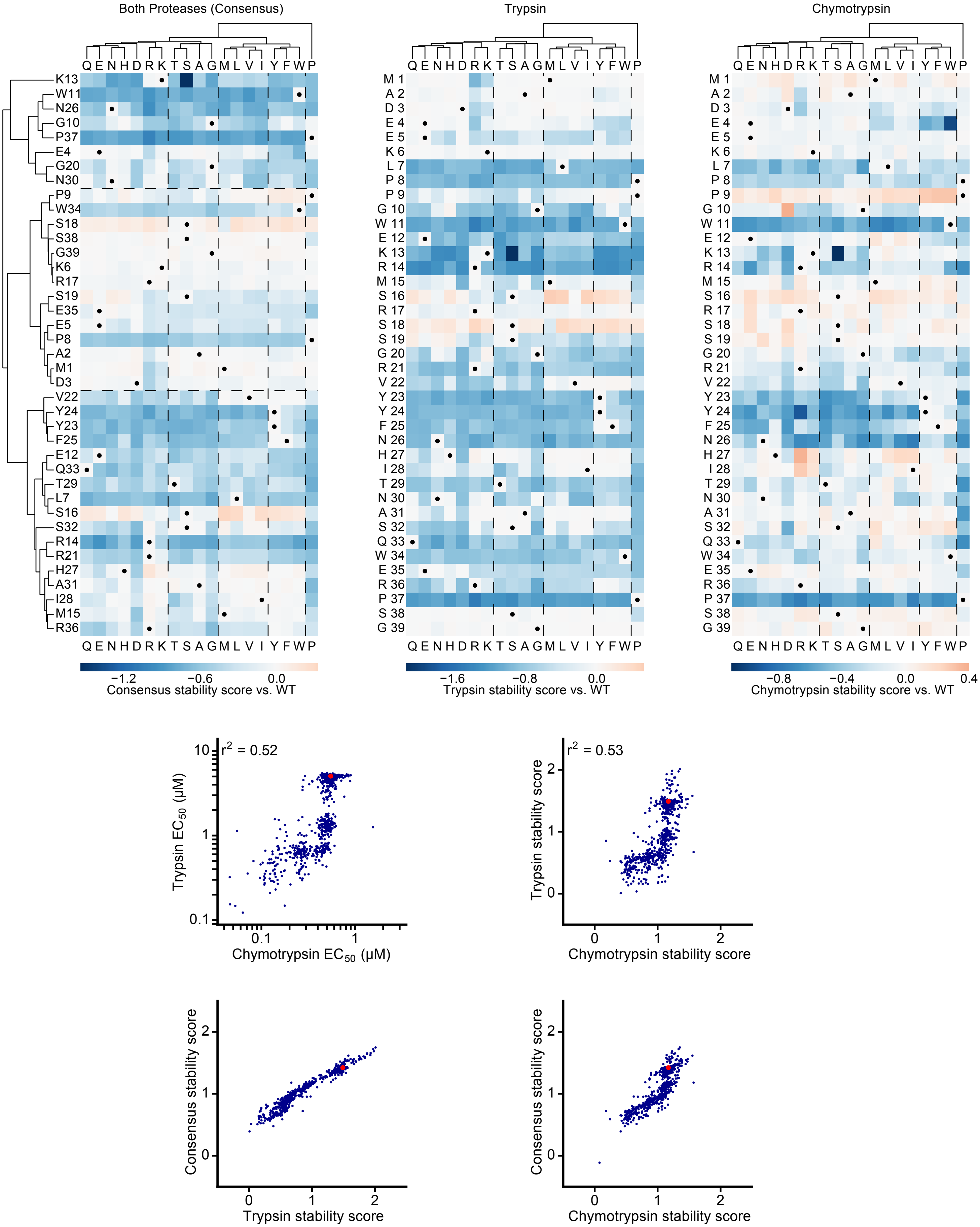


Fig. S10 Q: villin HP35

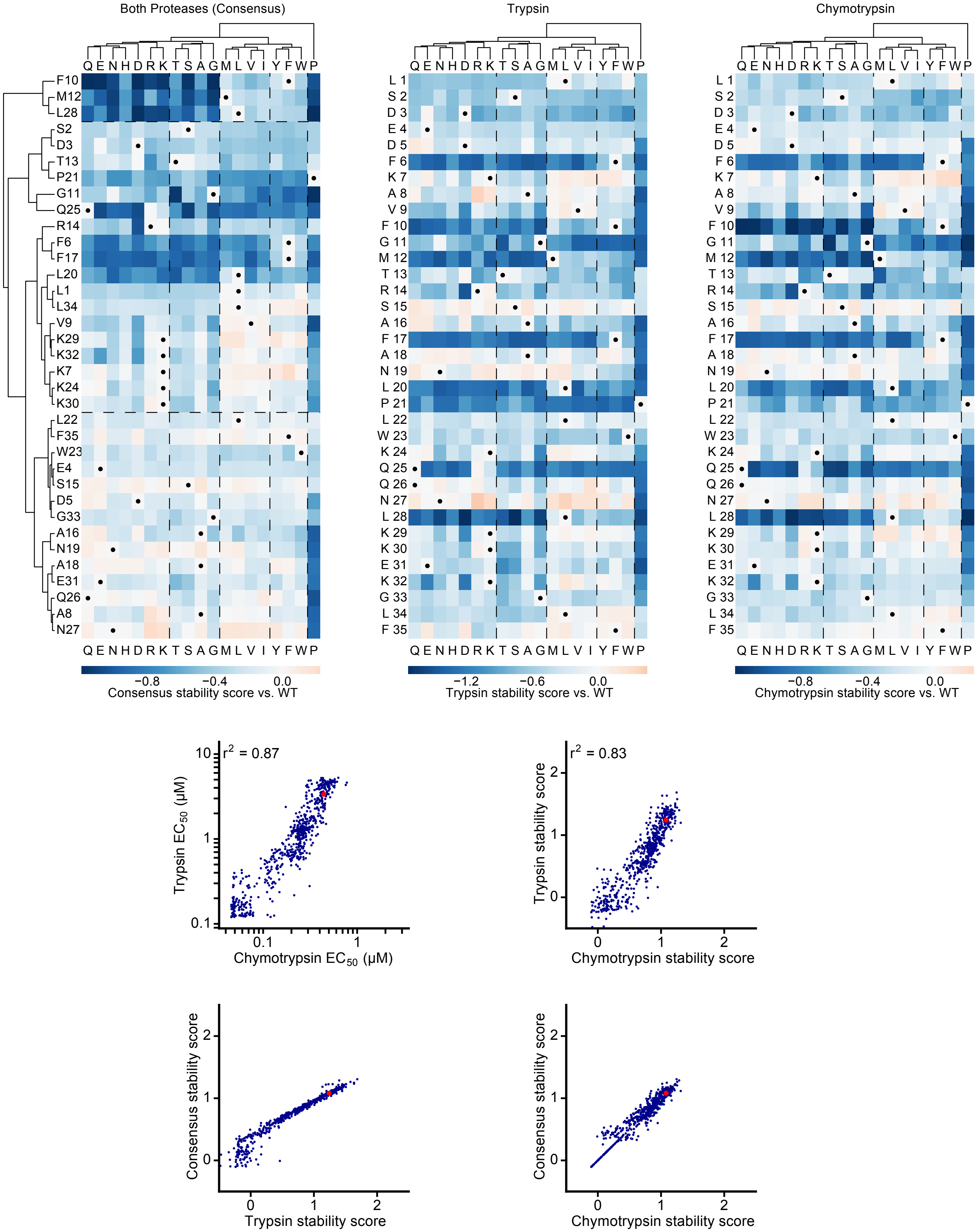


Fig. S10. Stability of all point-mutants in fourteen designed and three naturally occurring proteins. Top: For each protein analyzed by saturation mutagenesis, we show the consensus, trypsin, and chymotrypsin stability scores for all mutants in heat-map format. The wild-type amino acid at each position is indicated by a black dot. To cluster amino acids (shown at top), we represented each of the 19 amino acids (excluding Cys) as a vector of stability scores with one score per position for all positions in the 17 proteins examined. These vectors were used to compute the similarity between different amino acids by the Euclidian distance formula, and the amino acids were then clustered using this distance matrix and average linkage hierarchical clustering. To cluster positions in a given structure (shown at left for the consensus heat map), we represented each position as a vector containing all pairwise stability score differences between amino acids at that position. This representation causes positions to be clustered by their absolute amino acid preferences, rather than by the optimality of the wild-type amino acid (i.e. whether mutations are stabilizing or destabilizing). We used the Euclidian distance formula to compute similarity between positions, and positions were clustered using Ward's linkage hierarchical clustering. Dashed lines indicating column and row clusters are drawn as an aid to the eye. Residue numbering for villin HP35 begins at residue 1 for the first residue in HP35 (residue 42 in the numbering in Figs. 4B and Fig. S12). **Bottom:** For the set of mutants of a given wild-type protein, the consistency between trypsin and chymotrypsin results is shown using four scatter plots. (Upper left) Consistency between chymotrypsin and trypsin EC_{50} measurements. (Upper right) Consistency between chymotrypsin and trypsin stability scores. (Lower left) Consistency between trypsin stability scores and the derived consensus stability scores. (Lower right) Consistency between chymotrypsin stability scores and the derived consensus stability scores. (**A-N)** Stability profiles of all mutants in designed proteins. These profiles were assessed for agreement with their designed structures by examining the patterns of conservation in helices, sheets, and loops. All designs were qualitatively consistent with their designed structures except EEHEE_rd3_1702 (M). The profile for this protein does not demonstrate regular secondary structure, such as the pattern of buried and exposed strand residues exhibited by the other EEHEE designs (L and N).

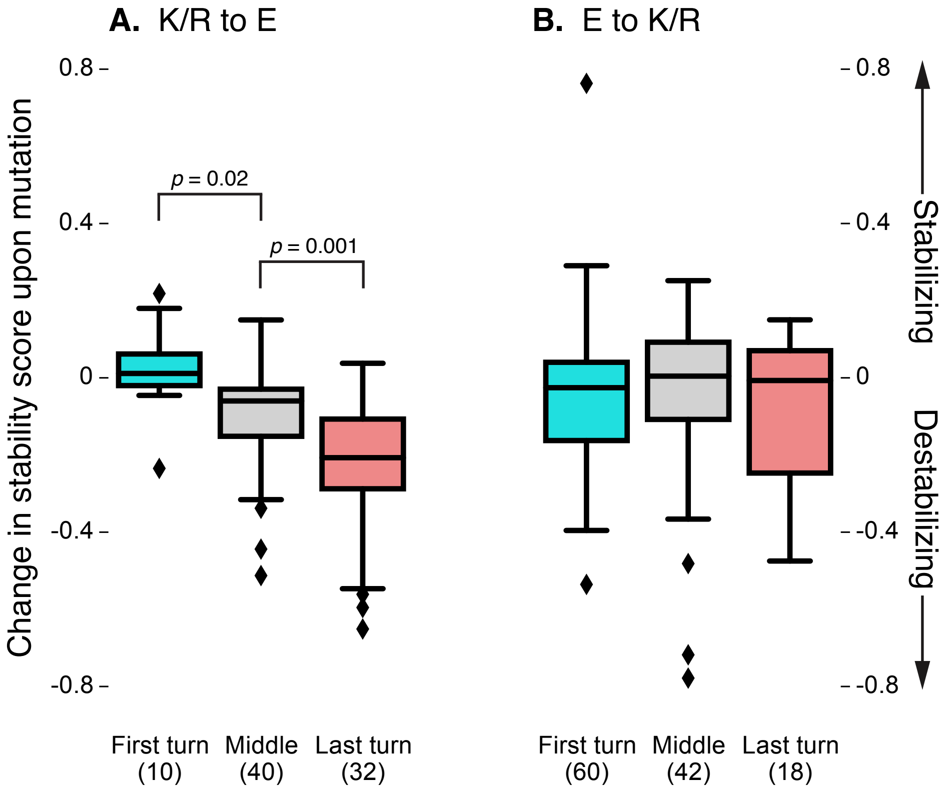


Fig. S11. Stability effects of charge-reversal mutations at first and last helical turns. In Fig. 4G and H, we showed that first and last helical turns have stronger preferences for particular charges compared with helical sites in general. However, because these first and last turns are already heavily populated with favorable charges in the stable designs, this effect could be attributed to a general tendency of the designed structures to stabilize the original, designed amino acid compared with a mutant amino acid (because all the original amino acids in the vicinity were designed to form mutually stabilizing interactions). To control for this effect, we compared charge-reversal mutations at first turn, middle, and last turn positions that all were designed with the same charge, to ensure that bias favoring the starting residue would be equally present at all three types of positions. Aspartate positions and mutations were excluded from this analysis due to Asp's unique favorability as a helix N-cap, which is distinct from the interactions made by the other charged residues. Each box-plot shows the distribution of changes in stability score for charge reversal mutations at first turn, middle, and last turn positions, with the total number of mutations examined shown in parentheses. The significance of the differences in mean was evaluated using the unequal variance (Welch's) *t*-test. (**A**) Positive-to-negative mutations (K or R to E) were neutral-to-favorable at first helical turns, mildly destabilizing at middle helical sites, and unfavorable at final helical turns. Positive-to-negative mutations at first helical turns were 0.12 ± 0.04 (mean ± SEM) stability score units more favorable at first helical turns compared with middle helical sites, and 0.13 ± 0.04 stability score units less favorable at final helical turns compared with middle helical sites. (**B**) However, no effect was observed for negative-to-positive (E to R or K) mutations at sites in the designs that were originally designed to be Glu.

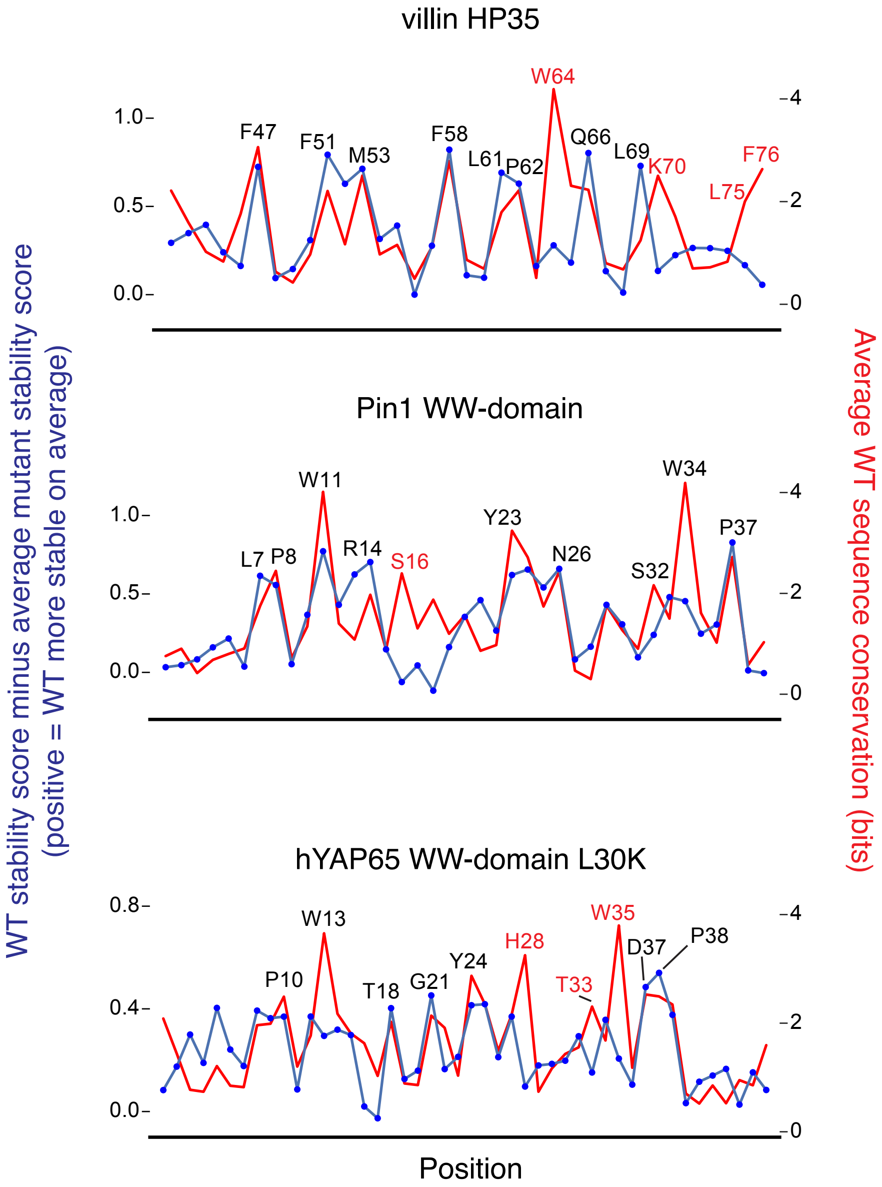


Fig. S12. Stability versus conservation in native proteins. For each protein, the average difference in stability between the wild-type sequence and all possible point mutant sequences for a given position is plotted in blue (left y-axis, positive values indicate the wild-type amino acid is more stable that the average mutant at that position). The level of conservation at each position is plotted in red (right y-axis, a larger number of bits indicates greater conservation). Key conserved residues are identified by name, and residue names highlighted in red indicate conserved positions where the wild-type residue is either destabilizing or neutral compared to possible mutants, indicating positions that are conserved for functional reasons. The consensus stability score for the mutant residue positions was used as the stability measure; see *Methods: Mutational stability effects*, Fig. S9, and Fig. S10 O (hYAP65), P (pin1), and Q (villin).

Table S1. Purification and characterization of selected designs.

Protein	Expression method	Melting temp. (°C)	ΔG_{unf}, 25°C (kcal/mol)	Trypsin stability score	Chymotrypsin stability score
HHH_rd1_0005	pET-28b+	Partially folded, not reversible		2.14	1.03
HHH_rd1_0092	pET-28b+	71	1.02	2.01	1.31
HHH_rd1_0142	pET-28b+[1]	84	2.81	2.34	1.52
HHH_rd1_0320	pET-28b+	84		1.90	1.57
EHEE_rd1_0284	SUMO	>95	4.70	1.62	1.51
EHEE_rd1_0407	SUMO	72	3.50	2.14	1.34
HHH_rd2_0134	pET-28b+	>90	4.53	1.88	1.43
EHEE_rd2_0005	pET-28b+	>95	No upper baseline	1.91	1.92
EHEE_rd2_0303	SUMO	76	1.94	1.06	0.98
HEEH_rd2_0127	SUMO	Poor stability, no lower baseline		1.19	1.20
HEEH_rd2_0771	pET-28b+	48	No lower baseline	1.09	1.20
HEEH_rd2_0779	SUMO	43	1.22	1.31	1.08
EEHEE_rd2_0770	pET-28b+	88		1.37	1.28
HHH_rd3_0006	pET-28b+	85	1.70	1.84	1.30
HHH_rd3_0008	pET-28b+	>90	4.34	2.29	1.90
HEEH_rd3_0223	SUMO	56		1.81	1.41
EEHEE_rd3_1049	pET-28b+	82	3.18	1.87	1.73
HEEH_rd4_0049	pET-28b+	65		1.72	1.30
HEEH_rd4_0053	pET-28b+	No upper baseline		1.84	1.38
HEEH_rd4_0094	pET-28b+	71	1.78	1.31	1.38
HEEH_rd4_0097	pET-28b+	85	2.65	1.51	0.91
HEEH_rd4_0349	pET-28b+	76	1.84	1.46	0.81

1. CD spectra in Fig. 3B-D and Fig. S6A from protein purified with pET-28b+; the NMR structure in Fig. 3A from protein purified using the SUMO system.

Table S2. Summary of data and refinement statistics for the NMR-derived structures.

Design ID	HHH rd1 0142	EHEE rd1 0284	HEEH rd4 0097	EEHEE rd3 1049
PDB ID (BMRB code)	5UOI	5UP5	5UYO	5UP1
NMR constraints				
Total NOEs	406	263	645	457
Intra-residual	133	99	177	169
Inter-residual	273	164	468	288
Sequential (i – j = 1)	104	77	175	117
Medium-range (1 < i – j < 5)	74	33	145	56
Long-range (i – j ≥ 5)	95	54	148	115
Hydrogen Bonds	0	8	0	6
Dihedral Angles				
ϕ	0	19	0	23
ψ	0	19	0	22
Structural Statistics				
Violations				
Distance constraints (Å)	0.0096 ± 0.0022	0.012 ± 0.0063	0.0117 ± 0.0029	0.010 ± 0.004
Dihedral angle constraints (°)	N/A	0.87 ± 0.30	N/A	0.99 ± 0.32
Max. distance constraint violation (Å)	< 0.30	< 0.30	0.35	< 0.30
Max. dihedral constraint violation (°)	N/A	6.3	N/A	7.1
Deviations from idealized geometry				
Bond lengths (Å)	0.014 ± 0.0004	0.014 ± 0.0003	0.0143 ± 0.0003	0.014 ± 0.0003
Bond angles (°)	0.92 ± 0.035	0.95 ± 0.027	0.899 ± 0.03	0.89 ± 0.023
Impropers (°)	1.70 ± 0.17	1.87 ± 0.13	1.70 ± 0.15	1.73 ± 0.15
Ramachandran plot[a,b]				
Most favored (%)	92.6	84.2	91.8	89.4
Additionally allowed (%)	7.4	15.6	8.1	10.4
Generously allowed (%)	0	0.2	0.1	0.2
Disallowed (%)	0	0	0	0
Average pairwise r.m.s.d. (Å)[c]				
Heavy	1.5	1.5	1.2	1.5
Backbone	0.8	0.8	0.5	0.8
Structure Quality Factors (raw/Z-scores)[d]				
Procheck G-factor (phi/psi)	-0.01/0.28	-0.58/-1.97	-0.03/0.20	-0.35/-1.06
Procheck G-factor (all)	-0.02/-0.12	-0.27/-1.60	-0.09/-0.53	-0.14/-0.83
Verify 3D	0.49/0.48	0.41/-0.80	-0.40/0.96	0.25/-3.37
MolProbity clashscore	12.29/-0.58	11.03/-0.37	15.14/-1.07	8.92/-0.01

[a] Based on Procheck analysis (*83*)

[b,c] Calculated for ordered regions. HHH_rd1_0142 (residues 4-13, 18-28 and 32-38), EHEE_rd1_0284 (residues 3-6, 9-20, 26-29 and 35-39), HEEH_rd4_0097 (residues 22-63 - note that residue 22 (numbered 3 in the PDB file) is the first designed residue due to the presence of a HIS tag and cleavage sequence that was retained in the experiment), EEHEE_rd3_1049 (residues 22-25, 30-33, 36-49, 55-57 and 60-62 - again, residue 22 (numbered 22 in the PDB file) is again the first designed residue).

[d] Obtained using the Protein Structure Validation Software (PSVS) Suite (*84*), made freely available at: *http://psvs-1_5-dev.nesg.org*

Table S3. Coefficients of logistic regressions in Fig. 2I and Fig. S8. Input data were standardized by subtracting the mean and dividing by the standard deviation; this standard deviation for each term is given here as the *scale* factor. The magnitudes of the resulting *weights* represent the relative importance of each standardized term in the regression.

Note: the following metrics were used in model fitting (all other metrics were excluded): abego_res_profile, abego_res_profile_penalty, avg_all_frags, avg_best_frag, buried_np_AFILMVWY_per_res, buried_np_per_res, cavity_volume, contact_all, contig_not_hp_avg, contig_not_hp_max, degree, exposed_hydrophobics, exposed_np_AFILMVWY, exposed_polars, exposed_total, fa_atr_per_res, fa_elec, fa_rep_per_res, fa_sol, frac_helix, frac_sheet, hbond_bb_sc, hbond_lr_bb, hbond_lr_bb_per_sheet, hbond_sc, hbond_sr_bb, hbond_sr_bb_per_helix, holes, hphob_sc_contacts, hydrophobicity, mismatch_probability, n_charged, n_hphob_clusters, net_atr_net_sol_per_res, net_atr_per_res, net_sol_per_res, netcharge, omega, p_aa_pp, pack, ref, score_per_res, ss_sc, unsat_hbond, worst6frags, worstfrag

In addition to the above score terms, the following metrics were used for the models for the EHEE topology:
abd50_mean, abd50_min, dsc50_mean, dsc50_min, ssc50_mean, ssc50_min

a. HHH topology

term	scale	weight
buried_np_AFILMVWY_per_res	6.143938	1.106409
avg_all_frags	0.173062	-0.420379
net_atr_net_sol_per_res	0.180124	-0.40754
n_charged	2.605848	0.368823
score_per_res	0.138613	-0.301226
abego_res_profile_penalty	0.008087	0.226585
fa_atr_per_res	0.144013	-0.19036
ss_sc	0.034898	0.179962
p_aa_pp	1.030702	-0.173117
avg_best_frag	0.054539	-0.159811
hbond_lr_bb	0.567897	0.154852
hphob_sc_contacts	3.787565	0.138973
pack	0.059774	-0.128618
mismatch_probability	0.030397	-0.110828
unsat_hbond	1.363714	-0.083367
contact_all	35.436644	-0.077172
hbond_sr_bb	1.515739	-0.023483
hbond_sc	3.528416	-0.023145
exposed_polars	130.426374	0.000487

b. EHEE topology

term	scale	weight
buried_np_AFILMVWY_per_res	5.373743	0.693898
ssc50_mean	0.218088	0.431708
contact_all	33.519194	0.403505
avg_all_frags	0.210121	-0.371572
ref	9.173606	0.363076
exposed_polars	114.822804	0.297397
p_aa_pp	1.540825	-0.255101
n_charged	3.006345	0.251671
net_atr_per_res	0.137426	-0.206979
exposed_np_AFILMVWY	77.497695	0.166332
ss_sc	0.039497	-0.122782
hbond_sc	4.198789	-0.115573
dsc50_min	0.019637	0.094124
omega	1.096478	0.077276
dsc50_mean	0.021069	0.077072
abego_res_profile_penalty	0.015392	0.070056
hphob_sc_contacts	3.130241	-0.06364
contig_not_hp_max	3.055984	0.059211
unsat_hbond	1.175985	-0.057431
hbond_bb_sc	1.567202	0.030324
degree	0.172861	0.021928
fa_atr_per_res	0.153013	-0.019776

c. HEEH topology

term	**scale**	**weight**
exposed_np_AFILMVWY	108.935585	0.427274
net_atr_net_sol_per_res	0.190478	-0.398769
n_charged	3.668114	0.387325
net_atr_per_res	0.15294	-0.223726
unsat_hbond	1.470219	-0.207302
degree	0.296797	-0.182265
hphob_sc_contacts	3.438678	0.161213
buried_np_per_res	3.939645	0.128862
avg_best_frag	0.107656	-0.122454
exposed_total	155.040818	0.106261
frac_sheet	0.037793	-0.092732
mismatch_probability	0.065516	-0.059097
hbond_lr_bb	1.246011	0.032511
abego_res_profile	0.050155	0.032083
p_aa_pp	1.705632	-0.029404
contact_all	39.650542	0.01471

d. EEHEE topology

term	scale	weight
exposed_total	124.25595	0.422062
buried_np_AFILMVWY_per_res	3.516422	0.417319
score_per_res	0.11383	-0.406039
avg_all_frags	0.150939	-0.33675
contig_not_hp_avg	0.422811	-0.29729
hbond_sr_bb	1.420808	-0.260861
contact_all	31.966	0.24149
buried_np_per_res	2.742073	0.166274
p_aa_pp	1.286554	-0.139461
hphob_sc_contacts	2.87758	0.116834
mismatch_probability	0.037403	-0.11664
exposed_np_AFILMVWY	79.642089	0.092047
holes	0.723974	0.077161
n_charged	1.907525	0.076617
hbond_sr_bb_per_helix	0.07735	-0.070124
avg_best_frag	0.033579	-0.06469
ss_sc	0.032323	0.040672
degree	0.218716	-0.040553
unsat_hbond	1.226997	0.037326
cavity_volume	4.774545	0.028378
hbond_lr_bb_per_sheet	0.079494	-0.026369
omega	0.977409	-0.020569
net_sol_per_res	0.12525	-0.020208
pack	0.050995	0.015764
hbond_lr_bb	1.174016	-0.014027
hbond_sc	2.806526	0.00671

e. All topologies

term	scale	weight
avg_all_frags	0.36899	-1.155357
net_atr_net_sol_per_res	0.228365	-0.851858
n_charged	4.039123	0.737556
buried_np_AFILMVWY_per_res	7.156103	0.687571
avg_best_frag	0.122679	-0.505259
fa_atr_per_res	0.297286	-0.457628
exposed_polars	143.838986	0.380774
unsat_hbond	1.52519	-0.335897
mismatch_probability	0.111063	-0.333612
hbond_lr_bb	6.08303	-0.323919
exposed_np_AFILMVWY	105.821904	0.323463
fa_rep_per_res	0.062033	0.256859
degree	0.322779	-0.237492
p_aa_pp	1.958557	-0.223329
netcharge	2.836751	-0.134041
worstfrag	0.454171	-0.109936
frac_sheet	0.154928	0.099065
buried_np_per_res	4.086264	0.099038
abego_res_profile_penalty	0.017412	0.086598
hbond_sc	5.015579	-0.063582
holes	1.102023	0.058121
cavity_volume	8.002496	-0.053517
score_per_res	0.182268	-0.048758
hydrophobicity	276.038472	0.047034
hbond_bb_sc	2.313826	0.033924
ss_sc	0.045131	0.030373
contig_not_hp_max	3.448408	0.021977
contact_all	41.227203	0.021457
omega	1.0144	0.018656
exposed_hydrophobics	110.978994	-0.01496
contig_not_hp_avg	1.081984	-0.000575

Definition of scoring metrics

Simple sequence and topological properties:

description: the design name
sequence: the design sequence
dssp: the design secondary structure, according to the DSSP algorithm (*85*)
n_res: the number of residues in the design
nres_helix: the number of helical residues in the design, according to DSSP
nres_sheet: the number of beta strand residues in the design, according to DSSP
nres_loop: the number of loop residues in the design, according to DSSP
frac_helix: nres_helix / n_res
frac_sheet: nres_sheet / n_res
frac_loop: nres_loop / n_res
n_charged: the count of D, E, K, and R residues in the designed sequence, plus one-half the number of H residues.
netcharge: the net charge on the design, assuming a charge of +1 on R and K, +0.5 on H, and -1 on D and E.
AlaCount: the count of Ala residues in the design
n_hydrophobic: the count of A, F, I, L, M, V, W, and Y residues in the design
n_hydrophobic_noA: the count of F, I, L, M, V, W, and Y residues in the design

Rosetta energy terms:

dslf_fa13, fa_atr, fa_dun, fa_elec, fa_intra_rep, fa_intra_sol_xover4, fa_rep, fa_sol, hbond_bb_sc, hbond_lr_bb, hbond_sc, hbond_sr_bb, lk_ball_wtd, omega, p_aa_pp, pro_close, rama_prepro, ref, ss_sc, total_score, yhh_planarity: all the scores in the Rosetta full-atom energy function. See (*55*) for documentation.

Simple combinations of Rosetta energy terms:

score_per_res: total_score / n_res
fa_atr_per_res: fa_atr / n_res
fa_rep_per_res: fa_rep / n_res
hbond_lr_bb_per_res: hbond_lr_bb / n_res
hbond_lr_bb_per_sheet: hbond_lr_bb / nres_sheet
hbond_sr_bb_per_helix: hbond_sr_bb / nres_helix
net_atr_per_res: (fa_atr + fa_rep) / n_res
net_sol_per_res: (fa_sol + fa_elec) / n_res
net_atr_net_sol_per_res: net_atr_per_res + net_sol_per_res

Rosetta filters:

See https://www.rosettacommons.org/docs/latest/scripting_documentation/RosettaScripts/Filters/Filters-RosettaScripts for all documentation.

cavity_volume: void volume inside the designed structure, in Å^3, computed with CavityVolume filter

degree: average number of residues in a 9.5 Å sphere around each residue, computed with AverageDegree filter

contact_all: number of sidechain carbon-carbon contacts in the designed structure, computed with AtomicContactCount filter

exposed_hydrophobics: exposed nonpolar surface area of the designed structure, in Å^2, computed using TotalSasa filter, set to compute hydrophobic-only SASA

exposed_polars: exposed polar surface area of the designed structure, in Å^2, computed using TotalSasa filter, set to compute polar-only SASA

exposed_total: total exposed surface area of the designed structure, in Å^2, computed using TotalSasa filter

fxn_exposed_is_np: exposed_hydrophobics / exposed_total

holes: a normalized measure of the void volume inside the designed structure, computed with Holes filter

helix_sc: the average shape complementarity of each helical secondary structure element with the rest of the structure, computed using SSShapeComplementarity filter, set to evaluate helices

loop_sc: the average shape complementarity of each loop element with the rest of the structure, computed using SSShapeComplementarity filter, set to evaluate loops only

mismatch_probability: the geometric average probability (across all positions in the design) that the designed residues will *not* adopt their designed secondary structures, as calculated by the PSIPRED algorithm (*75*) from the designed sequence. Computed using the SSPrediction filter.

pack: a normalized measure of packing density, computed using PackStat filter

unsat_hbond: number of buried, unsatisfied hydrogen bonding atoms, computing using the unsat_hbond filter

ss_sc: the average shape complementarity of each helical or loop element with the rest of the structure, computed using SSShapeComplementarity filter

BuriedUnsatHbonds filter

unsat_hbond2: number of buried, unsatisfied hydrogen bonding atoms, computed using BuriedUnsatHbonds2 filter

Custom metrics computed in Rosetta:

These metrics are not built-in Rosetta filters, but are computed within the Rosetta software

buried_np: buried nonpolar surface area in the designed structure on all amino acids, computed using version1 definitions of total nonpolar surface area per residue

buried_np_per_res: buried_np / n_res

buried_minus_exposed: buried_np - exposed_hydrophobics

buried_np_AFILMVWY: buried nonpolar surface area in the designed structure on nonpolar amino acids (AFILMVWY), computed using version2 definitions of total nonpolar surface area per residue

buried_np_AFILMVWY_per_res: buried_np_AFILMVWY / n_res

buried_over_exposed: buried_np / exposed_hydrophobics

exposed_np_AFILMVWY: exposed nonpolar surface area in the designed structure on nonpolar amino acids (AFILMVWY)

one_core_each: the fraction of secondary structure elements (helices and strands) with one large hydrophobic residue (FILMVYW) at a position in the core layer of the designed structure

two_core_each: the fraction of secondary structure elements (helices and strands) with two large hydrophobic residues (FILMVYW) at positions in the core layer of the designed structure

ss_contributes_core: the fraction of secondary structure elements (helices and strands) with one large hydrophobic residue (FILMVYW) at a position in either the core or interface layer of the designed structure

res_count_core_SASA: the number of residues in the core layer of the designed structure, with layers defined using solvent accessible surface area-based criteria

res_count_core_SCN: the number of residues in the core layer of the designed structure, with layers defined using sidechain neighbors-based criteria

percent_core_SASA: res_count_core_SASA / n_res

percent_core_SCN: res_count_core_SCN / n_res

Custom metrics computed using external scripts:

abego_res_profile: Each position *i* in the designed structure can be classified by its ABEGO type (*86*), and the ABEGO types of positions *i*-1, *i*, and *i*+1 form a triad that defines the three-residue local structure at a coarse level. The abego_res_profile metric is the sum over all positions *i* in the designed structure of log ((p_aa | abego triad) / (p_aa)), where (p_aa | abego triad) is the frequency of the designed amino acid (from position *i*) in regions of natural proteins sharing the same ABEGO triad as the designed region centered on position *i*, and p_aa is the overall frequency of the designed amino acid at position *i*. At each position, this score is positive when the designed amino acid is overrepresented (compared with its normal frequency) in regions of natural proteins with the same local ABEGO triad structure as the designed region, and the score is negative when the designed amino acid is underrepresented in regions of natural proteins with the same local ABEGO triad structure.

abego_res_profile_penalty: Same as abego_res_profile, except summing over only positions with negative abego_res_profile scores (positions where the designed residue is typically underrepresented in the local structure).
contig_not_hp_avg: average size of the contiguous (in primary sequence) regions of the designed sequence lacking a large hydrophobic residue (FILMVWY)
contig_not_hp_norm: contig_not_hp_avg / (n_res / (1 + n_hydrophobic_noA))
contig_not_hp_max: the size of the largest contiguous region (in primary sequence) in the designed sequence containing no large hydrophobic residues (FILMVWY)
contig_not_hp_internal_max: the size of the largest contiguous region (in primary sequence) in the designed sequence containing no large hydrophobic residues (FILMVWY), excluding the regions between the first and last large hydrophobic residues and the termini
hphob_sc_contacts: the total number of sidechain-sidechain contacts between large hydrophobic residues (FILMVWY) in the designed structure
hphob_sc_degree: hphob_sc_contacts / n_hydrophobic_noA
largest_hphob_cluster: the size of the largest group of large hydrophobic residues (FILMVWY) that are all connected by at least one contact to each other in the designed structure
n_hphob_clusters: the number of disconnected groups of large hydrophobic residues (FILMVWY), where a group is defined as residues that contact each other in the designed structure but do not contact residues outside of the group
hydrophobicity: total hydrophobicity of the designed sequence, using the amino acid hydrophobicity scale from (*87*)

Protease cut site counts:
chymo_cut_sites: the number of F, Y, W residues in the design
chymo_with_LM_cut_sites: the number of F, L, M, Y, W residues in the design
tryp_cut_sites: the number of K, R residues in the design
nearest_chymo_cut_site_to_Cterm: number of residues between the design C-terminus and the nearest F, Y, or W residue. If no F, Y, or W present, then the length of the design.
nearest_chymo_cut_site_to_Nterm: as previously, except to the design N-terminus.
nearest_chymo_cut_site_to_term: the minimum of nearest_chymo_cut_site_to_Nterm and nearest_chymo_cut_site_to_Cterm
nearest_tryp_cut_site_to_Cterm, nearest_tryp_cut_site_to_Nterm, nearest_tryp_cut_site_to_term: as defined for nearest_chymo_cut_site, except referring to K and R residues.

Helix end analysis:
These metrics indicate the number of charged residues and the net favorable charge at the first and last turns of helices. The first turn is defined as the first three helical residues based on the designed DSSP string, and the last turn is defined as the last three helical residues in the DSSP string. Amino acids D and E contribute one favorable charge in first turns and one unfavorable

charge in last turns, and K and R contribute one unfavorable charge in first turns and one favorable charge in last turns, with H as one-half the strength of K and R.
T1_netq: the net favorable charge in the first turn of each helix, summed over all helices.
T1_absq: the number of charged residues in the first turn of each helix, summed over all helices.
Tminus1_netq: the net favorable charge in the last turn of each helix, summed over all helices.
Tminus1_absq: the number of charged residues in the last turn of each helix, summed over all helices.
Tend_netq: T1_netq + Tminus1_netq
Tend_absq: T1_absq + Tminus1_absq

Fragment quality analysis:
Fragments were chosen for each designed protein using the standard Rosetta fragment generation protocol (*17*), which uses the designed sequence and PSIPRED-predicted secondary structure (*75*) as input. These metrics quantify the geometric agreement between the selected 9-mer fragments from natural proteins and the corresponding 9-mer segments of the designs (200 9-mer fragments are chosen per designed segment).
avg_all_frags: the average RMSD of all selected fragments to their corresponding segments of the designs, in Å. (200 × (*n* - 8) fragments in total for protein length *n*)
avg_best_frags: the average RMSD of the lowest-RMSD fragment for each designed segment, in Å. (*n* - 8 fragments in total)
sum_best_frags: the sum of the RMSDs of the lowest-RMSD fragment for each designed segment. (*n* - 8 fragments in total)
worstfrag: among the set of fragments that are the lowest-RMSD fragments for their positions, the highest RMSD found
worst6frags: among the set of fragments that are the lowest-RMSD fragments for their positions, the sum of the RMSDs of the six highest RMSD fragments

Tertiary motif analysis:
See (*88*) for all documentation. Positive values are favorable for all scores.
abd50_mean: the average of the tertiary motif abundance score across all residues in the design
abd50_min: the minimum tertiary motif abundance score across for any residue in the design
dsc50_mean: the average of the tertiary motif design score across all residues in the design
dsc50_min: the minimum tertiary motif design score across for any residue in the design
ssc50_mean: the average of the tertiary motif structure score across all residues in the design
ssc50_min: the minimum tertiary motif structure score across for any residue in the design

Regression scores for selecting Round 4 designs (round 4 only):
linear_reg_pred: the predicted stability score for each Round 4 design, according to the topology-specific linear regression models, parameterized on the results from design rounds 1-3

logistic_reg_pred: the probability that each Round 4 design will have a stability score above 1.0, according to the topology-specific logistic regression models, parameterized on the results from design rounds 1-3. Coefficients of the models are given in Table S3 A-D.

gb_reg_pred: the predicted stability score for each Round 4 design, according to the gradient boosting regression model, parameterized on the results from design rounds 1-3

logistic_alltop_pred: the probability that each Round 4 design will have a stability score above 1.0, according to the overall logistic regression model parameterized on results from all topologies together from design rounds 1-3. Coefficients of the model are given in Table S3E.

Explanation of included datasets

Eight supporting datasets are included with this work:
counts_and_ec50s.tar.gz
stability_scores.tar.gz
design_pdbs.tar.gz
design_scripts.tar.gz
design_structural_metrics.tar.gz
fig1_thermodynamic_data.csv
protein_and_dna_sequences.tar.gz
unfolded_state_model_params

These are annotated below.

counts_and_ec50s.tar.gz contains the raw deep sequencing counts from each proteolysis experiment, the EC_{50} value inferred for each sequence from these data according to the selection model, the credible interval for each EC_{50} value according to the selection model, and comparisons between the observed counts and the counts predicted by the selection model using the fitted parameters for all sequences. These data are contained in the *.fulloutput files. The file experiments.csv contains summary information about the different selections. EC_{50} values and other other associated columns (ec50_95ci_lbound, ec50_95ci_ubound, ec50_95ci) are given in units of $\log_3$ protease concentration (rounds 1-4) or $\log_2$ protease concentration (ssm2). An EC_{50} value of "1" corresponds to the lowest protease concentration tested in that experiment; a value of 2 corresponds to a 3-fold (or 2-fold) higher concentration, and so forth (see *Methods: Yeast display proteolysis* for a complete list of protease concentrations).

Column headers for *.fulloutput are as follows. These files are whitespace separated.
name: The name of the sequence for that row. Designs are named according to the topology (HHH, EHEE, HEEH, or EEHEE), the design round (rd1-4), and a number identifying each design. Names ending in ".pdb_random" refer to fully scrambled versions of a designed sequence (round 1 only). Names ending in ".pdb_hp" refer to scrambled versions of designed sequences preserving the hydrophobic or polar character at each position (round 1 only). Names ending in ".pdb_PG_hp" refer to scrambled versions of designed sequences preserving the hydrophobic or polar character at each position as well as the locations of all prolines and glycines (rounds 2-4). Names ending in ".pdb_buryD" refer to the control sequences where a single buried aspartate residue was inserted into each designed sequence (rounds 2-4).
counts0-13: Number of deep sequencing reads matching each sequence at the protein level (in the case of rounds 1-4) or at the nucleotide level (in the case of the site-saturation mutagenesis library "ssm2"). Each column 0-13 refers to a different selection condition within the experiment. Most experiments have only seven selection conditions (unselected and six protease concentrations, for columns 0-6). The round 2 and round 3 experiments have 14 selection

conditions because data from both replicates is combined and analyzed together. The selection conditions are defined in experiments.csv.
downsamp_counts0-13: Estimated number of cells collected for each sequence at each selection condition
pred_counts0-13: Predicted number of cells collected for each sequence at each selection condition according to the parameterized selection model
delta_llh0-13: The difference in log-likelihood between the predicted number of cells collected (pred_counts) and the actual number of cells collected (downsamp_counts) for each sequence at each selection condition, according to the parameterized selection model
signed_delta_llh0-13: Same as delta_llh, except multiplied by 1 or -1 so that positive values indicate that the parameterized model predicts a larger number of cells collected than were observed, and negative values indicate that the parameterized model predicts a smaller number of cells collected than were observed.
sel_k: the value of K used in model fitting, fixed at 0.8 for all analysis
sum_delta_llh: the sum of the delta_llh0-13 columns, a measure of goodness of fit
sum_signed_delta_llh: the num of the signed_delta_llh0-13 columns, a measure of goodness of fit
ec50_95ci_lbound: the lower boundary of the 95% credible interval of the inferred EC_{50} value
ec50_95ci_ubound: the upper boundary of the 95% credible interval of the inferred EC_{50} value
ec50_95ci: the size of the 95% credible interval of the inferred EC_{50} value
ec50: the inferred EC_{50} value
EC_{50} values are given in units of $\log_3$ protease concentration (rounds 1-4) or $\log_2$ protease concentration (ssm2). An EC_{50} value of "1" corresponds to the lowest protease concentration tested in that experiment; a value of 2 corresponds to a 3-fold (or 2-fold) higher concentration, and so forth (see *Methods: Yeast display proteolysis* for a complete list of protease concentrations).

The selection conditions and summary results for each round of each selection are defined in experiments.csv. The column headers are:
input: the name of the set of selection experiments (named by round and by protease)
column: the name of the column in the relevant *.fulloutput file
parent: the number of the library (row) that serves as the input library for the given selection
selection_strength: the concentration of protease applied to the given selection. Protease concentrations are given in $\log_3$ units (rounds 1-4) or $\log_2$ units (ssm2). A protease concentration of "1" corresponds to the lowest protease concentration used in that experiment, as listed in *Methods: Yeast display proteolysis*.
conc_factor: notes whether protease concentrations increased by a factor of 3 or a factor of 2 from experiment to experiment
parent_expression: the fraction of cells (events) passing the selection threshold in the given library before proteolysis, according to the sorting instrument

fraction_collected: the fraction of cells (events) passing the selection threshold in the given library after proteolysis, according to the sorting instrument
matching_sequences: the fraction of deep sequencing reads matching a sequence of interest for this library
cells_collected: the total number of cells (events) collected during the given selection, according to the sorting instrument

stability_scores.tar.gz contains the stability score data for all sequences, as computed from each sequence's EC_{50} value and predicted EC_{50} value in the unfolded state. Data is separated by design library (rd1-4 and the saturation mutagenesis library labelled ssm2) and both trypsin and chymotrypsin data is included in each file. Columns referring to trypsin results are labeled "_t"; columns referring to chymotrypsin results are labelled "_c". EC_{50} values and other other associated columns (ec50_95ci_lbound, ec50_95ci_ubound, ec50_95ci, delta_ec50, ec50_pred, delta_pred_vs_wt, ec50_rise) are given in units of $\log_3$ protease concentration (rounds 1-4) or $\log_2$ protease concentration (ssm2). An EC_{50} value of "1" corresponds to the lowest protease concentration tested in that experiment; a value of 2 corresponds to a 3-fold (or 2-fold) higher concentration, and so forth (see *Methods: Yeast display proteolysis* for a complete list of protease concentrations). Column headers for these files are as follows. These files are whitespace separated.

name: The name of the sequence for that row. Designs are named according to the topology (HHH, EHEE, HEEH, or EEHEE), the design round (rd1-4), and a number identifying each design. Names ending in ".pdb_random" refer to fully scrambled versions of a designed sequence (round 1 only). Names ending in ".pdb_hp" refer to scrambled versions of designed sequences preserving the hydrophobic or polar character at each position (round 1 only). Names ending in ".pdb_PG_hp" refer to scrambled versions of designed sequences preserving the hydrophobic or polar character at each position as well as the locations of all prolines and glycines (rounds 2-4). Names ending in ".pdb_buryD" refer to the control sequences where a single buried aspartate residue was inserted into each designed sequence (rounds 2-4).
sequence: The designed sequence as ordered for the oligo library. Includes padding sequences out to 43 residues in length (rds. 1-3, ssm2) or 50 residues in length (rd4) as described in *Methods: DNA synthesis*.
my_wt (ssm2 only): The name of the wild-type protein for a given mutant sequence
pos (ssm2 only): The position of the mutation for a given sequence, set to 0 for wild-type sequences (no mutations).
mut (ssm2 only): The amino acid that was mutated into the present sequence, set to "na" for wild-type sequences
wt_aa (ssm2 only): The wild-type amino acid that was changed in the present sequence, set to "wt" for wild-type sequences

assay_library (rds. 1-4 only): Identical for all sequences in each library, this column simply notes the DNA library (round 1, 2, 3, or 4) described in each file.
ec50 (_t, _c): the inferred EC_{50} value
delta_ec50 (_t, _c) (ssm2 only): ec50 - $ec50_{wt}$, where $ec50_{wt}$ is the EC_{50} of the wild-type version of a given mutant
ec50_95ci_lbound (_t, _c): the lower boundary of the 95% credible interval of the inferred EC_{50} value
ec50_95ci_ubound (_t, _c): the upper boundary of the 95% credible interval of the inferred EC_{50} value
ec50_95ci (_t, _c): the size of the 95% credible interval of the inferred EC_{50} value
ec50_pred (_t, _c): the predicted EC_{50} value for each sequence in its unfolded state, according to the unfolded state model. Note that these have different values for the same sequences in the ssm2 table versus the other tables because of (1) differences in scale ($\log_2$ units vs. $\log_3$ units) and (2) differences in location (different lowest protease concentrations were used for each set of assays).
delta_pred_vs_wt (_t, _c) (ssm2 only): ec50_pred - ec50_pred_{wt}
ec50_rise (_t, _c): ec50 - ec50_pred
stabilityscore (_t, _c): The stability score for each sequence, in units of $\log_{10}$ [protease]. Calculated as log_{10} ($base^{ec50\_rise}$), where *base* is 3 for the designed libraries (rd1 to rd4) and 2 for the ssm2 library.
stabilityscore (rds. 1-4 only): the minimum of the trypsin stability score (stabilityscore_t) and the chymotrypsin stability score (stabilityscore_c), used for overall ranking of designs, control sequences, and natural protein sequences in Fig. 2 and Fig. 5.
ec50_rise_c_adj (ssm2 only): Same as ec50_rise_c, except adjusted ("adj") so that all mutant unfolded state predicted ec50 values for chymotrypsin are no more than 1.4 ec50 units (on same scale as ec50_pred for the ssm2 experiment, which is $\log_2$ [protease]). This adjustment was made because the chymotrypsin unfolded state model appeared overly sensitive on the mutant library data (see *Methods: Unfolded state model*)
stabilityscore_c_adj (ssm2 only): Same as stabilityscore_c, except computed using ec50_rise_c_adj instead of ec50_rise_c.
consensus_ec50_rise (ssm2 only): The consensus ec50_rise for the given mutant, based on the protein-specific linear relation between trypsin and chymotrypsin stability scores (i.e. the unique best-fit line between trypsin and chymotrypsin stability scores for all mutants of the same wild-type protein). in units of $\log_2$ [protease]. See *Methods: Mutational stability effects*.
consensus_stability_score (ssm2 only): The consensus stability score for the given mutant on a log10 scale, calculated as log_{10} ($2^{consensus\_c50\_rise}$). Used for all analysis in Fig. 4, Fig S11, and Fig. S12.

design_pdbs.tar.gz contains PDB files of all four rounds of designs.

design_scripts.tar.gz contains the Rosetta input files (.xml script files and other input parameters) used to generate the designs.

design_structural_metrics.tar.gz contains the properties of the designed models used for all design analysis and selection of designs for testing (e.g. Fig. 2, Table S2). These properties include Rosetta energies, Rosetta filter terms, and additional structural and sequence metrics calculated using custom scripts. There are two sets of files:
rd[1-4]_relax_scored_talaris2013.sc: all designs relaxed and scored with the "Talaris2013" version of Rosetta energy function; no other filter scores included.
rd[1-4]_relax_scored_beta_nov15.sc: all designs relaxed and scored with the "beta_nov15" version of the Rosetta energy function; including all other terms used in design analysis and selection.

The column headers are annotated in *Methods: Definition of scoring metrics*.

fig1_thermodynamic_data.csv tabulates the thermodynamic data shown in Fig. 1, as well as the EC_{50} values and stability scores measured here in high-throughput (also found in the complete Round 4 dataset). See Fig. 1 caption for references. The column headers are:
name: the mutant being examined
sequence: the 50 a.a. sequence being examined (all sequences were extended to 50 a.a. by random addition of N, G, and S residues)
ref: brief description of the reference where the data originated; see these references for how Tm and ΔG_{unf} were estimated from the raw data in each instance.
conditions: the experimental conditions for the Tm and ΔG_{unf} measurements
Tm: melting temperatures, in °C. These values are the y-axis for the hYAP65 data in Fig. 1.
deltaGunf thermal: the unfolding free energy ΔG_{unf} as calculated from thermal denaturation experiments, in kcal/mol. See "conditions" for the temperature used in the calculation. These values are the y-axis for the Pin1 and BBL data in Fig. 1.
deltaGunf chemical: the unfolding free energy ΔG_{unf} as calculated from chemical denaturation experiments. See "conditions" and the original references for experimental details. These values are the y-axis for the villin data in Fig. 1.
ec50_t: the measured trypsin EC_{50}, in units of $\log_3$ [trypsin]. A value of 1 corresponds to the lowest trypsin concentration tested (0.07 μM), a value of 2 corresponding to a 3-fold higher concentration, and so forth.
ec50_pred_t: the predicted trypsin EC50 for each sequence in the unfolded state, according to the unfolded state model, in the same units as ec50_t.
ec50_95ci_t: the width of the 95% credible interval of ec50_t according to the selection model, in the same units as ec50_t.
ec50_rise_t: equal to ec50_t - ec50_pred_t

stabilityscore_t: the trypsin stability scores, in units of $\log_{10}$ [trypsin] (all stability scores are in $\log_{10}$ units). Equal to $\log_{10}$ (3.0 ^ ec50_rise_t). In other words, the difference between the measured EC_{50} and the predicted EC_{50} in the unfolded state, on a $\log_{10}$ scale. These values are the x-axis for all trypsin data in Fig. 1.
ec50_c, ec50_pred_c, ec50_95ci_c, ec50_rise_c, and stabilityscore_c: analogous to the above, except specific to chymotrypsin. An ec50_c of 1 corresponds to the lowest chymotrypsin concentrated tested (0.08 µM), and each increment corresponds to a 3-fold higher concentration.

protein_and_dna_sequences.tar.tz contains the DNA sequences as ordered by oligo library synthesis. There is one table per library (five in total). The columns in each table are:
name: the unique name for each sequence.
protein_sequence: the protein sequence as reverse-translated for the preparation of oligo library DNA. Note that these sequences include padding residues to bring all sequences to the same length. For the exact designed sequence as modeled on the computer (with no padding residues), see the sequences in the **design_structural_metrics.tar.gz** tables.
coding_dna: the reverse-translated DNA encoding protein_sequence, computed by DNAWorks 2.0 (see *Methods: DNA synthesis*).
full_dna: the full-length oligo specified for oligo library synthesis, including adapters for homologous recombination and (in some cases) adapters for amplifying sub-pools out of the library.

unfolded_state_model_params contains the complete fitted parameters used for the unfolded state model. Along with the values shown in Fig. S3, this table includes the terms tot_l, ind_b, and max_sumweight (see *Methods: Unfolded state model* for the complete specification of the model). Note that the position-specific parameters given in this table differ from those shown in Fig. S3 by a constant value for each row - these constants were added to aid visualization and interpretability in Fig. S3. Adding a constant to all parameters in a row does not change the output of the model so long as the constant term c_0 in eqn. (19) is adjusted to compensate.